Kontexteffekte in Large-Scale Assessments

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

im Fach Psychologie

eingereicht an der
Lebenswissenschaftlichen Fakultät der
Humboldt-Universität zu Berlin

von Dipl.-Psych. Sebastian Weirich

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Richard Lucius

Gutachter/Gutachterin: 1. Prof. Dr. Oliver Lüdtke

2. Prof. Dr. Hans Anand Pant

3. Prof. Dr. Matthias Ziegler

Tag der Verteidigung: 13. Juli 2015

Inhaltsverzeichnis

Abbildungs- und Tabellenverzeichnis

Danksagung

Mein Dank gilt insbesondere Dr. Katrin Böhme, Prof. Dr. Oliver Lüdtke und Prof. Dr. Hans Anand Pant, die die Entstehung dieser Arbeit konstruktiv betreut, gefördert und unzählige wertvolle Hinweise und Anregungen für die Gestaltung des Manuskripts sowie die Darstellung der Analysen gegeben haben. Prof. Dr. Hans Anand Pant und Prof. Dr. Petra Stanat danke ich für die Möglichkeit, am Institut zur Qualitätsentwicklung im Bildungswesen promovieren zu dürfen. Prof. Dr. Manuel Völkle, Prof. Dr. Oliver Lüdtke, Prof. Dr. Hans Anand Pant, Prof. Dr. Matthias Ziegler und Dr. Gizem Hülür danke ich für ihre Bereitschaft zur Begutachtung der Dissertation. Martin Hecht und Alexander Robitzsch danke ich für die vielen unermüdlichen Diskussionen und hilfreiche Ratschläge, die mein Verständnis für latente logistische Messmodelle signifikant gefördert haben. Allen Koautoren der in dieser Dissertation gebündelten Einzelbeiträge danke ich besonders für ihre angenehme und konstruktive Kooperation bei der Erstellung der Manuskripte. Dr. Heino Reimers möchte ich für das akribische Korrekturlesen des Manuskripts danken. Nicht zuletzt möchte ich mich bei den Kolleginnen und Kollegen des Instituts zur Qualitätsentwicklung im Bildungswesen für die angenehme und produktive Zusammenarbeit in den letzten Jahren bedanken.

Zusammenfassung

Kontexteffekte stellen im Rahmen von Large-Scale Assessments ein besonderes Problem dar, da sie dazu führen können, dass die aus Modellen der Item Response Theorie (IRT) abgeleiteten Parameter in substanzieller Weise verfälscht sind. Kontexteffekte können, müssen aber nicht diese Konsequenzen haben. Um in Einzelfällen abschätzen zu können, ob Kontexteffekte auftreten und dadurch die Gefahr verzerrter Parameter gegeben ist (und falls ja, in welcher Weise), müssen IRT-Modelle entwickelt werden, die zusätzlich zu Item- und Personeneffekten Kontexteffekte parametrisieren.

Diese Dissertation setzt sich aus drei Einzelbeiträgen zusammen. In dem ersten Beitrag wird ein solches Modell zur Schätzung von Kontexteffekten vorgestellt. Dabei werden Positionseffekte als ein Beispiel für Kontexteffekte in einem allgemeinen linearen gemischten Modell untersucht, und es werden die statistischen Eigenschaften dieses Modells im Rahmen einer Simulationsstudie evaluiert. Hier zeigt sich vor allem die Bedeutung des Testdesigns: nicht nur, um trotz Kontexteffekten zu validen Parameterschätzern im Raschmodell zu gelangen, sondern um auch die statistischen Gütekriterien komplexerer IRT-Modelle zur Modellierung von Kontexteffekten zu gewährleisten.

Der zweite Einzelbeitrag bezieht zusätzlich zu Kontexteffekten auf der Itemseite Kontexteffekte auf der Personenseite mit ein. Dabei wurde eine signifikante Interaktion aus Positions- und Motivationseffekten gefunden.

Der dritte Beitrag befasst sich mit dem Problem fehlender Werte auf Hintergrundvariablen in Large-Scale Assessments. Als Kontexteffekt wird in diesem Beispiel derjenige Effekt verstanden, der die Wahrscheinlichkeit eines fehlenden Wertes auf einer bestimmten Variablen systematisch beeinflusst. Dabei wurde das Prinzip der multiplen Imputation auf das Problem fehlender Werte auf Hintergrundvariablen übertragen. Anders als bisher praktizierte Ansätze (Dummy-Codierung fehlender Werte) konnten so in einer Simulationsstudie für fast alle Simulationsbedingungen unverfälschte Parameter auf der Personenseite gefunden werden.

Ein Schwerpunkt dieser Dissertation liegt dabei auf dem Problem der Identifizierung und der Balancierung von Kontexteffekten, sowie der wechselseitigen Abhängigkeit beider Prozesse.

Schlüsselbegriffe: Large-Scale Assessments, Item-Response-Theorie, Kontexteffekte, Testdesign, Itempositionseffekte, missing data, multiple Imputation

Abstract

Context effects are considered as a particular problem in large-scale assessments, which may lead to substantially biased parameters in IRT analyses. Context effects may, but do not necessarily have those consequences. To decide whether context effects occur in individual cases and lead to biased parameters, specific IRT models have to be developed which parametrize context effects additionally to item and person effects.

The present doctoral thesis consists of three single contributions. In the first contribution, a model for the estimation of context effects in an IRT framework is introduced. Item position effects are examined as an example of context effects in the framework of generalized linear mixed models. Using simulation studies, the statistical properties of the model are investigated, which emphasizes the relevance of an appropriate test design. A balanced incomplete test design is necessary not only to obtain valid item parameters in the Rasch model, but to guarantee for unbiased estimation of position effects in more complex IRT models.

The second contribution additionally examines context effects on the person side. In a cross-level interaction model, the interaction of position effects and test-taking effort was significant and substantial.

The third contribution deals with the problem of missing background data in large-scale assessments. The effect which predicts the probability of a missing value on a certain variable, is considered as a context effect. Statistical methods of multiple imputation were brought up to the problem of missing background data in large-scale assessments. In contrast to other approaches used so far in practice (dummy coding of missing values) unbiased population and subpopulation estimates were received in a simulation study for most conditions.

The focus of the present doctoral thesis lies on the problem of identifying and balancing context effects and the interdependency of both.

Keywords: Large-scale assessments, item response theory, context effects, test design, item position effects, missing data, multiple imputation

Liste der Beiträge

Die Online-Version der Dissertation beinhaltet den zur Veröffentlichung eingereichten Beitrag (S. 73), während für die bereits veröffentlichten Beiträge die Quelle und der Link für die Artikel angegeben sind.

Erster Einzelbeitrag:

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*(7), 535-548. doi: 10.1177/0146621614534955

Zweiter Einzelbeitrag:

Weirich, S., Penk, C., Hecht, M., Roppelt, A., & Böhme, K. (under review). Item position effects are moderated by changes in test-taking effort. *Journal of Educational Measurement.*

Dritter Einzelbeitrag:

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education, 2*(9), 1-18.

(Stand: 10. August 2015)

“All models are wrong but some are useful” (Box & Draper, 1987; S. 74).

1. Überblick über die Dissertation

1.1 Struktur der Dissertation

Die vorliegende Arbeit befasst sich auf psychometrischer Ebene mit Problemen bei der modellbasierten Kompetenzmessung in Large-Scale Assessments. Sie setzt sich aus drei Einzelbeiträgen und einer Rahmung zusammen. In der Rahmung wird zunächst in einem Einleitungsteil (Kapitel 1.2) das Anliegen der Dissertation in einem inhaltlichen Überblick kurz umrissen, ohne dabei näher auf Fachbegriffe einzugehen. Zentrale Begriffe und Konzepte werden anschließend in den Kapiteln 2.1 bis 2.7 eingeführt. In Kapitel 3 erfolgt eine kurze Zusammenfassung des ersten Einzelbeitrags, bevor in einem Zwischenkapitel (Kapitel 4) offen gebliebene Fragestellungen des ersten Einzelbeitrags eingehender diskutiert werden. Kapitel 5 dient als Einführung für den zweiten Einzelbeitrag, der in Kapitel 6 kurz zusammengefasst wird. Kapitel 7 dient wiederum als Einführung für den dritten Einzelbeitrag, der in Kapitel 8 zusammengefasst wird. Daran schließt sich eine Gesamtdiskussion der Befunde sowie ein abschließendes Fazit an.

Der Anhang dieser Arbeit enthält die drei Einzelbeiträge sowie gegebenenfalls Anhänge zu den jeweiligen Artikeln („online appendix“).

1.2 Inhaltlicher Überblick über die Dissertation

Large-Scale Assessments kommen in der empirischen Bildungsforschung international und seit der letzten Dekade auch national verstärkt zum Einsatz. Sie verfolgen das Ziel, Kompetenzen von Schülerinnen und Schülern in objektiver, reliabler und valider Weise zu messen. Die dabei verwendeten Verfahren sind in mehrfacher Hinsicht komplex; sowohl was die verwendeten Messmodelle, die Testdesigns sowie die Bedingungen der Testadministration betrifft. Insbesondere die Messmodelle sind an strikte mathematische oder statistische Annahmen geknüpft, die Voraussetzung für die Validität der in diesen Modellen geschätzten Parameter sind (Swaminathan, Hambleton & Rogers, 2007). Diese Annahmen können häufig nicht oder nur in eingeschränkter Weise erfüllt werden. Vor dem Hintergrund, dass die Ergebnisse

empirischer Bildungsforschung auf bildungsadministrativer Ebene einen hohen Stellenwert besitzen, ist es zwingend erforderlich, die Verlässlichkeit der Ergebnisse modellbasierten Kompetenzmessung in Large-Scale Assessments zu evaluieren. Konkret betrifft das folgende Fragen:

1. Inwieweit sind die den statistischen Messmodellen zugrundeliegenden theoretischen Annahmen in der empirischen Praxis von Large-Scale Assessments erfüllt? Können diese Annahmen gegebenenfalls über Bedingungen der Testadministration oder die Wahl eines geeigneten Testdesigns hergestellt werden?
2. Wenn diese theoretischen Annahmen nicht oder nicht vollständig gewährleistet werden können: Was sind die möglichen (unerwünschten) Auswirkungen auf die Güte der Messung? Können diese unerwünschten Auswirkungen minimiert werden, und wenn ja, mit welchen Maßnahmen?
3. Welche Methoden oder Verfahren erlauben es, für eine konkrete Testsituation zu entscheiden, ob die gewonnenen Ergebnisse valide sind?

Die drei Einzelbeiträge dieser Dissertation werden sich den oben stehenden Fragen nur exemplarisch nähern können und sie anhand eines konkreten Einzelbeispiels – etwa einer bestimmten, nicht gewährleisteten theoretischen Modellannahme – betrachten.

Im Kapitel 2 der Rahmung soll die modellbasierte Kompetenzmessung in Large-Scale Assessments eingeführt werden: Was bedeutet „modellbasiert“ und welche Besonderheiten sind in Large-Scale Assessments zu berücksichtigen? Ferner werden in Kapitel 2.3 Kontexteffekte eingeführt und ihre möglichen Auswirkungen auf die Ergebnisse von Kompetenzmessungen beschrieben. Anhand eines prominenten Beispiels, der sogenannten *NAEP Reading Anomaly*, werden in Kapitel 2.5 die unerwünschten Konsequenzen, die aufgrund von Kontexteffekten resultieren können, aufgeführt. Aus dem Beispiel wird die Notwendigkeit der Modellierung von Kontexteffekten abgeleitet. Dies wiederum leitet zu dem ersten Einzelbeitrag „Modeling Item Position Effects Using Generalized Linear Mixed Models“ über, der die Modellierung eines spezifischen Kontexteffekts sowie Auswirkungen auf die Güte der geschätzten Modellparameter des Item-Response-Modells beschreibt. Der Beitrag behandelt ferner die Bedeutung von Testdesigns zur Kontrolle beziehungsweise zur Ausbalancierung unerwünschter Kontexteffekte.

Im Kapitel 4 der Rahmung wird auf ein Problem eingegangen, das im Diskussionsteil des ersten Einzelbeitrags zwar genannt, nicht aber gelöst wurde. Konkret geht es dabei darum, ob und warum Kontexteffekte trotz vollständig ausbalancierter Testdesigns zu verzerrten Parameterschätzungen führen können. Anhand eines der Anschaulichkeit zuliebe stark vereinfach-

ten Beispiels wird gezeigt, dass diese Verzerrung – zumindest für die hier betrachteten Positionseffekte – nur äußerst gering ist.

Der zweite Einzelbeitrag „Item Position Effects Moderated by Changes in Test-taking Effort“ behandelt die Frage, ob eine Annahme, die den meisten Testdesigns zugrunde liegt, Gültigkeit hat, nämlich dass Kontexteffekte von Effekten der Items oder der Personen unabhängig sind. Es wird gezeigt, dass dies zumindest für Positionseffekte nicht der Fall ist.

Hinführend zu dem dritten Einzelbeitrag, wird in Kapitel 7 näher auf die praktische Umsetzung von Large-Scale Assessments in der quantitativen empirischen Bildungsforschung eingegangen. Es wird beschrieben, dass die solchen Assessments zugrundeliegenden probabilistischen Modelle für eine konkrete Messung nicht eigentlich ein, sondern zwei Modelle in einer Abhängigkeitsstruktur umfassen, das Item-Response-Modell und das Populationsmodell, wobei die Parameter des zweiten (des Populationsmodells) bedingt auf die Parameter des ersten geschätzt werden. Um zu Parametern für die Fähigkeit von Personen zu gelangen, sind folglich mehrere Modellierungsschritte notwendig.

Der dritte Einzelbeitrag „Nested Multiple Imputation in Large-Scale Assessments“ behandelt mögliche Auswirkungen einer Kontextbedingung (hier der fehlenden Werte) auf das Populationsmodell. Anders als in den vorangegangenen Beiträgen wird hier nun das zweite der beiden Modelle betrachtet.

An die drei Einzelbeiträge schließt sich eine allgemeine Diskussion der Befunde an.

2. Modellbasierte Kompetenzmessung in Large-Scale Assessments

Das folgende Kapitel fasst in verkürzter Weise zentrale Aspekte modellbasierter Kompetenzmessung in Large-Scale Assessments zusammen. Dabei wird in Kapitel 2.1 das Messmodell beschrieben und darauf eingegangen, in welcher Weise Messmodelle fehlspezifiziert sein können. Anhand prominenter Literaturbeispiele wird in Kapitel 2.5 gezeigt, dass solche Fehlspezifizierungen gravierende Konsequenzen für die Güte der Itemparameterschätzung haben können, aber nicht müssen. Diese unerwünschten Konsequenzen können dabei explizit, also durch Anpassungen des Modells, oder implizit, das heißt durch Anpassungen des Testdesigns minimiert werden. Im ersten Fall wird versucht, ein möglichst korrekt spezifiziertes Messmodell zu formulieren. Dies bedeutet, dass das Modell um zusätzliche Parameter erweitert wird. Im zweiten Fall wird versucht, das Messmodell unverändert beizubehalten, auch wenn es gegebenenfalls fehlspezifiziert ist. Um dennoch unverfälschte Itemparameter zu gewinnen, können speziell konstruierte Testdesigns eingesetzt werden.

Ein Schwerpunkt der Dissertation ist dabei, die Wirksamkeit dieser zweiten Anpassungsmethode zu untersuchen: Unter welchen Umständen erlauben angepasste Testdesigns die Schätzung unverfälschter Itemparameter, selbst wenn das Messmodell fehlspezifiziert ist? Um diese Frage zu beantworten, müssen Kontexteffekte explizit modelliert werden. Diese Modellierung wird im ersten Einzelbeitrag der Dissertation näher behandelt.

2.1 Messmodelle

In der empirischen Bildungsforschung kommen Large-Scale Assessments in der Kompetenzdiagnostik, oder allgemeiner, Leistungsmessung zum Einsatz. Leistungsmessung, in Abgrenzung zur Messung von Einstellungen, fragt danach, wie fähig Personen in einer bestimmten Disziplin oder in einem bestimmten Kompetenzbereich sind, z. B. im Fach Mathematik oder im Kompetenzbereich „Lesen“ für das Fach Deutsch. Zentral ist dabei, dass diese zu messenden Merkmale oder Konstrukte als latent definiert sind. Sie können nicht direkt beobachtet werden (im Gegensatz etwa zur Größe einer Person), sondern müssen aus beobachteten Merkmalen *erschlossen* werden. In der Leistungsmessung werden diese beobachteten Merkmale als Variablen bezeichnet. Gemeint sind damit die Antworten der Personen auf Testaufgaben, genauer, ob die Person eine Aufgabe richtig oder falsch gelöst hat, beziehungsweise wie viele Punkte sie in einer Aufgabe erreicht hat. Die Messung latenter Merkmale erfordert daher immer ein spezifisches Messmodell, das definiert, in welcher Beziehung die beobachteten Variablen X zu dem latenten Konstrukt θ, dessen Messung von Interesse ist, stehen. Messmodelle werden aus Testtheorien abgeleitet.

Prinzipiell kommen für die Ableitung eines Messmodells zwei Testtheorien infrage, zum einen die Klassische Testtheorie (KTT), zum anderen die probabilistische oder Item Response Theory (IRT). Für eine *vergleichende* Kompetenzmessung, die die Leistung einer Population von Schülerinnen und Schülern auf einer kriterial definierten Skala abbilden will, eignet sich die IRT eher als die KTT, da Items und Personen auf einer gemeinsamen Skala abgebildet werden können, und die Schwierigkeit von Items direkt in Bezug zu der Fähigkeit von Personen gesetzt werden kann (Embretson & Reise, 2000; Hambleton, Swaminathan & Rogers, 1991). Swaminathan et al. (2007) führen aus, dass gerade diese Modellierung von Itemparametern der entscheidende Vorteil der IRT gegenüber der KTT ist. Aus der IRT wiederum lassen sich eine Vielzahl von Modellen ableiten, die eine Wahrscheinlichkeitsbeziehung zwischen dem latenten Merkmal θ, also der Fähigkeit der Person, und seinen manifesten Indikatoren, den beobachteten Variablen X, postulieren: Je höher die Merkmalsausprägung von θ,

desto höher ist die Wahrscheinlichkeit, dass diese Person eine Testaufgabe (ein Item *i*) korrekt lösen wird. Im einfachsten Modell der IRT, dem eindimensionalen dichotomen Raschmodell, kann man diese Relation linear darstellen:

$$\text{logit}\left(P(X_{ni}=1)\right)=\theta_n-\beta_i\,. \tag{1}$$

Die Wahrscheinlichkeit $P(X_{ni}=1)$, dass eine Person n das Item i löst, wird lediglich durch einen Personenparameter θ_n und einen Itemparameter β_i definiert, die in einer linearen Relation zueinander stehen. Da jedes Item durch nur einen Parameter beschrieben wird, spricht man beim Raschmodell auch vom einparametrischen oder 1PL-Modell. Aus Gleichung 1 geht hervor, dass eine Person mit der hypothetischen Fähigkeit 2 ein Item der Schwierigkeit 2 mit derselben Wahrscheinlichkeit lösen sollte, mit der eine Person der Fähigkeit 1 ein Item der Schwierigkeit 1 löst. Gleichung 1 besagt ferner, dass nicht die Wahrscheinlichkeit selbst durch θ_n und β_i vorhergesagt wird, sondern der Logit dieser Wahrscheinlichkeit. Das bedeutet, das Modell macht eine Vorhersage über einen transformierten Wahrscheinlichkeitswert. Die Transformation ist folgendermaßen definiert:

$$\text{logit}\left(P(X_{ni}=1)\right)=\ln\left(\frac{P(X_{ni}=1)}{1-P(X_{ni}=1)}\right). \tag{2}$$

Der Logit ist der natürliche Logarithmus des Wettquotienten. Der Wettquotient wiederum entspricht einer Wahrscheinlichkeit geteilt durch ihre Gegenwahrscheinlichkeit. Eine Wahrscheinlichkeit von 20% entspricht einem Logit von $\ln\left(\frac{0.2}{1-0.2}\right)\approx -1.39$, eine Wahrscheinlichkeit von 50% einem Logit von $\ln(1)=0$, und eine Wahrscheinlichkeit von 80% einem Logit von $\ln\left(\frac{0.8}{1-0.8}\right)\approx 1.39$. Die Transformation bewirkt, dass die auf einen Wertebereich von $[0,1]$ beschränkte Skala einer Wahrscheinlichkeit nun auf einen Bereich von $[-\infty,+\infty]$ transformiert wird, was besonders dann sinnvoll ist, wenn diese Wahrscheinlichkeit in linearen Regressionsmodellen durch andere Variablen vorhergesagt werden soll. Die Logit-Transformation ist dabei nur eine von mehreren Möglichkeiten, eine Transformation auf den Bereich von $[-\infty,+\infty]$ zu erhalten.

Ferner muss nun die Wahrscheinlichkeitsverteilung von X spezifiziert werden. Im Raschmodell ist sie folgendermaßen definiert:

$$X_{ni}\sim\text{binomial}\left(1,\pi_{ni}\right),\ \text{mit}\ \pi_{ni}=\left(P(X_{ni}=1)\right). \tag{3}$$

Die Gleichungen 1 und 2 definieren, wie groß *die durch das Modell vorhergesagte* Wahrscheinlichkeit π_{ni} ist, dass eine Person mit einer Fähigkeit θ_n ein Item mit der Schwierigkeit β_i

korrekt löst. Die konkrete Itemantwort (oder *item response*) X_{ni} wird dabei als eine Zufallsziehung aus einer Bernoulli-Verteilung mit dem Parameter π_{ni} definiert. Das Raschmodell besteht folglich aus den drei Komponenten (1) Modellgleichung, (2) Transformationsfunktion (*link function*) und (3) der Zufallskomponente (*random component*) (De Boeck et al., 2011).

Aus der Modellformulierung in Gleichung 1 folgt eine Reihe von Annahmen, deren empirische Gültigkeit Voraussetzung dafür sind, dass ein solches Modell die Item- und Personenparameter valide schätzt (vgl. Embretson & Reise, 2000).

1. Es gibt lediglich zwei Faktoren, θ_n und β_i, die die Lösungshäufigkeit eines Items systematisch beeinflussen. In der hier dargestellten Form bezeichnet θ_n dabei die eindimensionale Fähigkeit der Personen (*unidimensional latent trait*).
2. Die Items im Raschmodell unterscheiden sich lediglich in ihrer Schwierigkeit, nicht in ihrer Trennschärfe. Die Itemcharakteristik-Kurven (*item characteristic curves*, ICC) sind für alle Items parallel.
3. Die Items im Raschmodell sind lokal stochastisch unabhängig, das heisst, nach Kontrolle der eindimensionalen Personenfähigkeit θ_n existieren keine systematischen Beziehungen der Itemantworten zueinander mehr. In Gleichung 1 erkennt man das daran, dass die Wahrscheinlichkeit $P(X_{ni}=1)$ beispielsweise nicht noch *zusätzlich* davon abhängt, wie etwa das Item X_{i-1} beantwortet wurde.
4. Doppelte Monotonizität: Die Item-Response-Funktionen sind monoton steigend mit steigender Fähigkeit des Schülers. Bringt man für einen Schüler die Items in eine Reihenfolge nach ihrer erwarteten Lösungshäufigkeit, so ist diese Reihenfolge für jeden Schüler gleich.

Jenseits des Raschmodells existieren zahlreiche IRT-Modelle, die eine stärker differenzierte Vorhersage der Lösungswahrscheinlichkeit $P(X_{ni}=1)$ erlauben. Diese Modelle heben einige der oben beschriebenen Restriktionen auf und sollen hier vorerst nur in 4 Kategorien eingeteilt werden:

1. Modelle, die mehr als einen Parameter je Item schätzen: zwei-, drei- oder vierparametrische Modelle (2PL, 3PL, 4PL) schätzen für jedes Item zusätzlich zur Schwierigkeit einen Parameter für die Trennschärfe (2PL), für die Ratewahrscheinlichkeit (*guessing parameter*, 3PL) und die Flüchtigkeitswahrscheinlichkeit (*slipping parameter*, 4PL). Die ICCs von Items mit unterschiedlicher Trennschärfe sind nicht länger parallel; eine größere Trennschärfe bedeutet, dass die Kurve steiler verläuft. Eine Ratewahrscheinlichkeit größer Null bedeutet, dass die untere Asymptote der ICC größer als 0 ist, und

eine Flüchtigkeitswahrscheinlichkeit größer Null bedeutet, dass die obere Asymptote der ICC kleiner als 1 ist.

2. Modelle für polytome Daten erlauben die Modellierung von Daten, die nicht nur Unterscheidungen von 0 (falsch) und 1 (richtig) beinhalten, sondern in mehreren Kategorien beispielsweise von 0 (falsch), 1 (eher falsch), 2 (eher richtig), bis 3 (richtig) kodiert sind. Hierbei wird für jedes Item nicht eine ICC geschätzt, sondern $K-1$ ICCs, wobei K die Anzahl der Kategorien für dieses Item ist. Für ein Item mit 4 Kategorien (0, 1, 2, 3) würden demzufolge 3 ICCs geschätzt werden.
3. Mehrdimensionale IRT-Modelle definieren θ_n nicht als eindimensionales, sondern mehrdimensionales Konstrukt (McDonald, 2000; Tate, 2003).
4. Modelle, die über θ_n und β_i hinaus zusätzliche Effekte (oder Varianzkomponenten, z. B. Kontexteffekte) parametrisieren, können sinnvoll sein, wenn sich etwa im Raschmodell eine Verletzung der Annahme lokaler stochastischer Unabhängigkeit zeigt. Anders als in 2PL/3PL-Modellen wird hier nicht die Anzahl der Parameter je Faktor, sondern die Anzahl der Faktoren im Modell erweitert.

Praktische Anwendungen haben gezeigt, dass das Raschmodell mit seinen sehr strikten Annahmen empirische Daten häufig signifikant schlechter beschreibt, als es beispielsweise komplexere 2PL oder 3PL-Modelle vermögen (Hambleton et al., 1991). Ein solcher Befund für sich allein genommen ist erst einmal nicht überraschend, da im Falle genesteter Modelle das komplexere Modell (d. h. dasjenige mit mehr Parametern) die Daten gar nicht schlechter beschreiben kann. Es ist folglich eher sinnvoll, nach den praktischen Auswirkungen dieses „Modellmisfits“ im Raschmodell zu fragen (Messick, 1998; Sinharay & Haberman, 2014; Sinharay & Johnson, 2003; Swaminathan et al., 2007). Ähnlich argumentiert Zumbo (2007), für den die Frage der Validität von Testwerten oder Testmodellen bedeutet, „to consider the *consequences* of inferences from test scores“ (S. 50). Dafür wiederum muss sehr genau unterschieden werden, *was* bei der Messung von zentralem Interesse ist, also ob nach den praktischen Auswirkungen auf die Parameter der Items, der Personen, oder beidem gefragt wird.

Man könnte auch den umgekehrten Weg in Erwägung ziehen und im Zweifel einfach immer das komplexere Modell benutzen (da dieses ja empirisch die Daten mindestens genauso gut beschreibt wie das einfachere Modell). Dieses Vorgehen ist jedoch mit erheblichen Problemen bezüglich der Interpretation der Ergebnisse verbunden. So betonen Tuerlinckx und De Boeck (2004), dass in Modellen, die zusätzlich zu β_i und θ_n etwa lokale Abhängigkeiten zwischen den Itemantworten parametrisieren, die Itemparameter nicht länger die „natural and simple interpretation of the difficulty of an item” haben (S. 299). Der Itemparameter lässt sich

nicht länger direkt in Bezug zur Lösungswahrscheinlichkeit des Items setzen: „Second, the parameter β_2 does not have the natural interpretation of marking the point on the latent scale where the probability of a 1-response is .5" (S. 299). Tuerlinckx und De Boeck (2004) schließen: „[T]he parameters pertaining to a single item cannot be seen as item difficulties" (S. 307).

In verschiedenen praktischen Anwendungen wie etwa vergleichenden Schulleistungsstudien (Allen, Donoghue & Schoeps, 2001; Bonsen, Lintorf, Bos & Frey, 2008; Foy, Galia & Li, 2008; OECD, 2006, 2012; Pant et al., 2013; Stanat, Pant, Böhme & Richter, 2012) ist man jedoch gerade auf diese „natural and simple interpretation of the difficulty of an item" angewiesen, um die Ergebnisse in Bezug zu bereits etablierten Skalen oder Standards setzen zu können. Diese einfache Interpretierbarkeit ist also einerseits nötig, wenn man die Schwierigkeiten von Items direkt in Bezug zu den Fähigkeiten der Schülerinnen und Schüler setzen möchte, oder wenn man andererseits im Rahmen regelmäßiger (etwa jährlicher) Erhebungen eine Entwicklung oder Tendenz in der Leistung der Schülerpopulation abbilden will. Das Raschmodell besitzt als ein marginales Modell (Molenberghs & Verbeke, 2004; Wilson & De Boeck, 2004) Eigenschaften, die beides erlauben: Schwierigkeiten von Items und Fähigkeiten von Personen können auf einer gemeinsamen Skala abgebildet werden, und diese gemeinsame Skala kann als Vergleichsmaßstab oder Referenz zwischen verschiedenen disjunkten Personenkohorten genutzt werden (Embretson & Reise, 2000; Rost, 2004). Die sehr restriktiven Annahmen des Raschmodells sind so gesehen Fluch und Segen zugleich: Fluch, weil sie in empirischen Anwendungen meist nicht als gegeben gelten können, und Segen, weil diese Restriktionen Interpretationen erlauben, wie sie z. B. für Standard-Setting-Verfahren im Rahmen der Entwicklung von Kompetenzstufenmodellen (Cizek, 2001; Cizek & Bunch, 2007) für die Bildungsstandards (Granzer et al., 2009; Walther, Van den Heuvel-Panhuizen, Granzer & Köller, 2007) nützlich sind.

Wenn Large-Scale Assessments genutzt werden, um die Leistungen von Schülerinnen und Schülern in Relation zu metrisch und kriterial *bereits definierten* Skalen abzubilden, besteht demzufolge nicht die Möglichkeit, von dem zur Definition der bestehenden Skala verwendeten Messmodell, beispielsweise dem Raschmodell, abzuweichen. Zudem stellen komplexere Modelle deutlich höhere Anforderungen an die Stichprobengröße, um eine vergleichbare Präzision etwa für Schätzungen auf der Personenseite zu gewährleisten. Das zentrale Anliegen dieser Dissertation ist daher nicht zu zeigen, dass in empirischen Anwendungsfällen alternative Modelle die Daten oftmals besser beschreiben als das Raschmodell (das ist nahezu immer der Fall). Es soll vielmehr die Frage behandelt werden, welche Möglichkeiten existieren, aus

dem Raschmodell unverfälschte Parameter zu gewinnen, *obwohl* es in den meisten Anwendungsfällen im strengen Sinne ein fehlspezifiziertes Modell ist. Der Schwerpunkt liegt dabei nicht auf einer Korrektur des Modells oder einer nachträglichen Korrektur seiner geschätzten Parameter, sondern auf dem Testdesign. Die vorliegende Arbeit wird sich dabei auf den vierten Punkt beschränken, also Modelle betrachten, die zusätzliche Varianzkomponenten beinhalten. Solche zusätzlichen Effekte können etwa Kontexteffekte sein.

Hierzu wird in Kapitel 2.2 zunächst sehr kurz auf die Begriffe Unverfälschtheit und Validität eingegangen, anschließend sollen in Kapitel 2.3 zunächst Kontexteffekte kurz definiert und in 2.4 die formalen Implikationen solcher zusätzlicher Varianzkomponenten beschrieben werden. Nach einem Beispiel möglicher Konsequenzen der Nichtberücksichtigung von Kontexteffekten in Kapitel 2.5 wird in den Kapiteln 2.6 und 2.7 darauf eingegangen, welche Möglichkeiten seitens des Testdesigns existieren, diese Konsequenzen zu minimieren. Konkret soll es um die Frage gehen: Wenn Kontexteffekte auftreten, die durch das Raschmodell nicht parametrisiert werden, wie können dann aus dem Raschmodell dennoch valide Parameterschätzer gewonnen werden?

2.2 Unverfälschtheit und Validität

In der statistischen Literatur existieren verschiedene Definitionen zur Validität, wobei vorwiegend zwischen Inhalts-, Konstrukt- und Kriteriumsvalidität unterschieden wird (*content validity*, *construct validity*, *criterion-based validity*; Cronbach & Meehl, 1955; Messick, 1995; für einen Überblick, siehe Zumbo, 2007). In der Literatur zu latenten Messmodellen findet man darüber hinaus den Begriff der Validität des Messmodells (*measurement validity*; Zumbo, 2007), der oft synonym mit Unverfälschtheit (*unbiasedness*) gebraucht wird. Im Sinne Messicks (1984) ist ein latentes Messmodell valide, (a) wenn das durch die manifesten Indikatoren (Testitems) operationalisierte Konstrukt nicht unteridentifiziert (*underidentification*) ist, und (b) wenn die Testitems darüber hinaus keine konstrukt-irrelevante Varianz (*construct-irrelevant variance* oder *construct-irrelevant task difficulty*) beinhalten (Messick, 1984). Aus einem validen Messmodell resultieren valide oder unverfälschte Parameter (*unbiasedness*). In dieser Dissertation wird der Begriff Validität immer im Sinne von *measurement validity* gebraucht.

Zwei Beispiele für invalide Messmodelle (im Sinne nicht gegebener *measurement validity*) sollen kurz aufgeführt werden:

Angenommen ein Test zur Erfassung mathematischer Kompetenz würde lediglich Additions- und Subtraktionsaufgaben enthalten. Ein solcher Test würde das Konstrukt *Mathemati-*

sche Kompetenz nur unzureichend abbilden, weil er keine Aufgaben zum Multiplizieren, Dividieren etc. enthält. Demzufolge würde das Konstrukt *Mathematische Kompetenz* durch den Test nur unteridentifiziert abgebildet werden.

Wenn nun ein Test zur Erfassung mathematischer Kompetenz Sachaufgaben enthält, die sprachlich anspruchsvoll formuliert sind, ist es möglich, dass Personen, die prinzipiell die mathematische Kompetenz besitzen, die Aufgabe zu lösen, die Aufgabenstellung sprachlich nicht oder nicht vollständig erfassen können und deswegen an der Aufgabe scheitern. Die Aufgabe misst damit zusätzlich zur mathematischen auch noch einen Teil konstrukt-irrelevanter sprachlicher Kompetenz.

In dem folgenden Kapitel 2.3 wird ausgeführt, dass Kontexteffekte konstrukt-irrelevante Varianz in Messmodellen bedingen und folglich die *measurement validity* gefährden können.

2.3 Definition von Kontexteffekten

Brennan (1992) stellt fest, dass trotz umfangreicher Literatur bislang keine genaue Definition von Kontexteffekten existiert:

> „Another salient feature of the context-effects literature is a considerable lack of specificity about the definition of context effects. For example, the previously discussed four examples of context effects are typical in that the phrase ‚context effect' was never defined. […] Because context effects are seldom defined, the literature does not usually provide an unambiguous basis for judging whether or not a context effect exists. For example, from reading the context-effects literature, one might infer that many researchers would probably endorse the following statement: ‚A context effect exists whenever an examinee's response to an item in a test form is influenced by information included in other items in the same form.'" (Brennan, 1992; S. 235f).

Stewart (1981) beschreibt Kontexteffekte beispielsweise als „influences on test performance associated with the content of successively presented test items or sections". Bei Knowles (1988) sowie Khorramdel und Frebort (2011) werden statt einer Definition nur Beispiele für Kontexteffekte genannt, so etwa Positions- oder Reihenfolgeeffekte. Der „Kontext" eines Items sind dabei die anderen Items des Tests, beziehungsweise deren Eigenschaften, etwa ihre Schwierigkeit oder Reihenfolge. Obwohl also von *context effects* die Rede ist, sind damit im strengen Sinne lediglich Kontexteffekte auf der Itemseite (*item context effects*) gemeint (vgl. Ryan & Chiu, 2001). In dieselbe Richtung geht auch die Definition bei Wainer und Kiely (1987): „Context effects refer to any influence or interpretation that an item may

acquire purely as a result of its relationship to the other items making up a specific test" (S. 187).

Die dagegen von Brennan (1992) sowie Leary und Dorans (1985) aufgeführten Beispiele für Kontexteffekte (unterschiedliche Antwortformate, adaptive Tests, Benutzung von Taschenrechnern in Mathematiktests, Itemreihenfolge und -position) gehen über diese lediglich auf *item context effects* abzielende Definition hinaus und implizieren eine allgemeinere Begriffsbestimmung. So kann, wenn ein Teil der Testpersonen Taschenrechner benutzen darf, ein anderer jedoch nicht, dies als Kontexteffekt auf der Personenseite (*person context effect*) verstanden werden. Nach Brennan (1992) können Kontexteffekte „only through reference to … (a) a universe of generalization; and (b) a universe of allowable observations, which is a restricted version of the universe of generalization" (S. 236f) definiert werden. Es hängt demnach von dem "universe of generalization" ab, ob etwas als Kontexteffekt definiert werden kann oder nicht. Ein Beispiel soll das verdeutlichen:

Angenommen sei ein Test, der ausschließlich Items des Itemformats „multiple choice" enthält. Jedes Item hat dabei vier Antwortoptionen, von denen eine richtig ist. Für jedes Item wird dabei zufällig entschieden, ob die richtige Antwortoption an erster, zweiter, dritter oder vierter Stelle steht. Die Nummer der richtigen Antwortoption ist demzufolge eine *Eigenschaft des Items* und daher kein Kontexteffekt. Für ein einzelnes Item gibt es keine Variabilität an Möglichkeiten, also kein "universe of generalization". Die richtige Antwortoption wäre hier weder modellierbar noch sinnvoll als Kontexteffekt zu definieren. Wenn in einem zweiten Test jedoch diese Items in jeweils vier Varianten eingesetzt werden, so dass für jedes Item die richtige Antwortoption einmal an erster, einmal an zweiter, einmal an dritter und einmal an vierter Stelle steht, gibt es für jedes Item ein „Universum", das vier mögliche Bedingungen enthält. Die Position der richtigen Antwortoption könnte hier also als Kontexteffekt definiert und modelliert werden.

Das Beispiel ist analog auch auf Kontexteffekte auf der Personenseite übertragbar. So zählt etwa bei Perlini und Zumbo (1998) die Tageszeit der Testung zu Kontexteffekten. Wenn jedoch die Administrationsbedingungen des Tests so gestaltet sind, dass er für alle Testteilnehmer zur selben Uhrzeit stattfindet, würde das nicht länger als Kontexteffekt gelten können.

Angelehnt an Brennan (1992) sollen daher in dieser Arbeit unter Kontexteffekten alle nicht konstanten Einflüsse auf die gezeigte Testleistung verstanden werden, die nicht ausschließlich auf Eigenschaften der Person oder Eigenschaften der Items zurückgeführt werden können. Kontexteffekte können dabei auf der Item- und der Personenseite auftreten, sowie als Interaktion (*cross-level interaction*) auf beiden Seiten. Wenn etwa Mädchen in einem Test bessere

Leistungen erzielen als Jungen, wäre das kein Kontexteffekt auf der Personenseite, da das Geschlecht eine Eigenschaft der Person ist. Bearbeitet jedoch eine zufällig ausgewählte Hälfte der Testteilnehmer sämtliche Testaufgaben vormittags, die andere Hälfte nachmittags, und es finden sich im Mittel schlechtere Leistungen für die zweite Gruppe, wäre das als Kontexteffekt auf der Personenseite zu werten. Dass man in diesem Beispiel von einem Effekt „auf der Personenseite“ spricht, obwohl die Tageszeit doch gar keine Eigenschaft der Person ist, hat den Grund, dass der Faktor „Tageszeit“ zwischen Personen, und nicht zwischen Items variiert: Nicht die Items, sondern die Personen werden zufällig auf verschiedene Tageszeitbedingungen verteilt. Würde man dagegen sämtliche Kinder vormittags und nachmittags testen, dabei den gesamten normierten Test in zwei Hälften aufteilen (Testheft 1 und Testheft 2) und den Kindern vormittags Testheft 1 und nachmittags Testheft 2 vorlegen, und es würden sich im Mittel schlechtere Leistungen für Testheft 2 finden, so würde es sich bei der Tageszeit um einen Kontexteffekt auf der Itemseite handeln.

Abschließend soll ein Beispiel für *cross-level interaction* genannt werden. Dazu sollen wieder die beiden Faktoren „Tageszeit“ und „Testheft“ (Testheft 1 vs. Testheft 2) betrachtet werden. Die jeweils verwendeten Items seien normiert, haben also eine bekannte Schwierigkeit. Angenommen, die Items in Testheft 2 sind schwieriger als ihre Normierungswerte (beispielsweise weil die Schriftart in Testheft 2 sehr klein und schlecht lesbar ist). Dazu sei die Leistung der Kinder nachmittags schlechter als vormittags. Wenn nun die mittlere Schwierigkeitsdiskrepanz, d. h. die Abweichung der Itemschwierigkeiten im Testheft von ihren Normierungswerten, bei der Testdurchführung am Nachmittag größer ist als bei der Testdurchführung am Vormittag, hätte man es mit einer *cross-level interaction* zweier Kontexteffekte zu tun.

Theoretisch muss aber nicht jede Interaktion eine Interaktion zweier verschiedener Ebenen sein. Beispielsweise sind mehrere Kontexteffekte auf der Itemseite denkbar, die interagieren könnten, so etwa die Position eines Items im Testheft sowie die Position der richtigen Antwortoption des Items.

Formal können Kontexteffekte über zusätzliche Terme oder Varianzkomponenten im Messmodell (siehe Kapitel 2.1) parametrisiert werden. Diese zusätzlichen Terme erlauben es, die Varianz des Kontexteffekts von der Varianz der Items und/oder Personen zu trennen. Treten Kontexteffekte auf, ohne jedoch im Messmodell parametrisiert zu werden, können sie als eine Quelle konstrukt-irrelevanter Varianz die *measurement validity* (siehe Kapitel 2.2) gefährden.

2.4 Kontexteffekte auf der Itemseite: Verletzung lokaler stochastischer Unabhängigkeit

Lord und Novick (1968) führen aus, dass die Annahme lokaler stochastischer Unabhängigkeit im Raschmodell (vgl. Kapitel 2.1) äquivalent zur Annahme einer eindimensionalen latenten Fähigkeit θ_n (*unidimensional latent trait*) ist. Wenn keine weiteren Faktoren existieren, die über die eindimensionale Fähigkeit θ_n und β_i hinaus die Wahrscheinlichkeit der Itemantworten bedingen, müssen die Itemantworten – kontrolliert für θ_n – zueinander unabhängig sein (vgl. auch Swaminathan et al., 2007). Eine Möglichkeit, diese unterstellte Unabhängigkeit zu überprüfen, ist die sogenannte Q3-Statistik (Yen, 1984, 1993). Hier werden zunächst die Item- und Personenparameter des Messmodells unter Zugrundelegung dieser Unabhängigkeit spezifiziert. Anschließend werden für alle Itempaare die residualen Itemkorrelationen r_{ij} mit $i \neq j$ bestimmt, die, insofern die Unabhängigkeit tatsächlich gegeben ist, sämtlich 0 oder näherungsweise 0 betragen sollten. (Üblicherweise wird ein Toleranzbereich von $-.2 \leq r_{ij} \leq +.2$ zugelassen.) Liegt die Residualkorrelation für mehrere Itempaare außerhalb dieses Bereichs, könnte das einerseits bedeuten, dass es sich bei θ_n nicht um ein ein-, sondern ein mehrdimensionales Konstrukt handelt (Böhme & Robitzsch, 2009; Stout, 1987, 1990). Andererseits könnten auffällige Residualkorrelationen ein Indiz dafür sein, dass Faktoren oder Varianzkomponenten (beispielsweise Kontexteffekte), die die Wahrscheinlichkeit der Itemantworten bedingen, in dem zugrunde gelegten IRT-Modell nicht berücksichtigt worden sind (Monseur, Baye, Lafontaine & Quittre, 2011; Tuerlinckx & De Boeck, 2004). Das Umgekehrte gilt jedoch nicht: Unauffällige Residualkorrelationen sind keine Gewährleistung dafür, dass das zugrunde gelegte IRT-Modell alle Faktoren, die die Wahrscheinlichkeit der Itemantworten bedingen, auch beinhaltet (Sinharay & Haberman, 2014). Grund dafür ist zum einen, dass in der Q3-Statistik nur die (für θ_n kontrollierten) Kovarianzen berechnet werden. Diese könnten theoretisch auch dann Null sein, wenn die lokale stochastische Unabhängigkeit nicht gilt, die residualen Zusammenhänge jedoch beispielsweise quadratisch sind.

Verschiedene über das Testdesign realisierte Methoden der Ausbalancierung, die in den Kapitel 2.6 und 2.7 näher beschrieben werden, gewährleisten dagegen, dass die Kovarianzen 0 sind, obwohl im strengen Sinne keine stochastische Unabhängigkeit gegeben ist. Stout (1987) und Stout (1990) verwenden dafür den Begriff „essentielle Eindimensionalität“. Sie ist gegeben, wenn in einem Test die mittlere residuale Kovarianz über alle Paare von Testitems (kontrolliert für θ_n) klein ausfällt (Swaminathan et al., 2007). Stout, Nandakumar, Junker, Chang und Steidinger (1991) schlagen vor, Eindimensionalität zu prüfen, indem der Test in

zwei Hälften aufgeteilt wird, wobei eine Testhälfte ausschließlich Items enthält, deren eindimensionale Struktur bereits sichergestellt ist. Beide Testhälften sollten zu 1 korrelieren. Dieses Verfahren ist in dem Programm DIMTEST (Stout et al., 1991) implementiert.

Das Problem sowohl in DIMTEST als auch in der Q3-Statistik ist jedoch, dass keine Aussage darüber gemacht werden kann, ob die möglicherweise zutage tretende Verletzung der unterstellten dimensionalen Struktur zu verzerrten Parameterschätzern führt, und wenn ja, in welcher Größenordnung. Glas und Suarez Falcon (2003) konnten beispielsweise zeigen, dass die Item-Response-Kurven und Itemparameter trotz verletzter lokaler stochastischer Unabhängigkeit unverzerrt geschätzt werden konnten.

2.5 Beispiel: NAEP reading anomaly

Die Untersuchung von Glas und Suarez Falcon (2003) zeigt, dass Kontexteffekte nicht immer praktisch bedeutsame Auswirkungen haben müssen. Sie können jedoch durchaus ernsthafte Konsequenzen haben. Das bekannteste Beispiel dafür ist die sogenannte *NAEP reading anomaly*, die in einer amerikanischen Large-Scale Studie im Rahmen des *National Assessment of Educational Progress* (NAEP) auftrat (Beaton, 1988; Zwick, 1991). Beim Vergleich der Leseleistungen zweier altersgleicher Kohorten der Jahre 1984 und 1986 fand sich für 17-jährige Schülerinnen und Schüler eine unerwartet große Leistungsdifferenz: die Leseleistungen waren 1986 deutlich schlechter als 1984. In der darauffolgenden Erhebung im Jahr 1988 wurden in einer Zusatzstudie originale Testhefte aus den Jahren 1984 und 1986 abermals eingesetzt. In einem *common-population equating design* (Kolen & Brennan, 2004; von Davier, A., Carstensen & von Davier, 2008) konnten die vermeintlichen Leistungsunterschiede als „Artefakt“ einer abweichenden Zusammenstellung und einer abweichenden Position der Leseitems zwischen den Jahren 1984 und 1986 identifiziert werden. Diese unterschiedlichen Eigenschaften des Testinstruments wurden in dem 1986 zugrunde gelegten Messmodell nicht berücksichtigt. Die unterschiedlichen Kontextbedingungen zwischen 1984 und 1986 betrafen dabei sowohl die Position der Items als auch ihre Zusammenstellung: Die Leseitems waren 1984 im Testheft mit Schreibaufgaben, 1986 hingegen mit Mathematik- und Naturwissenschaftsaufgaben kombiniert. Aufgrund dieser, mehrere Kontextbedingungen gleichzeitig betreffenden Variation kann nicht genau quantifiziert werden, wie groß die Abweichung war, die durch Positionseffekte zustande kam, und wie groß die Abweichung war, die durch Effekte unterschiedlicher Aufgabenkombinationen zustande kam. Die *NAEP reading anomaly* gilt seither als Beleg für die möglichen praktischen Konsequenzen, die eintreten können, wenn Kontexteffekte

nicht adäquat berücksichtigt werden. Sie müssen daher in Large-Scale Assessments durch das Messmodell und/oder durch das Testdesign mit einbezogen werden.

In Kapitel 2.1 wurde ausgeführt, dass Kontexteffekte prinzipiell durch alternative IRT-Modelle, die zusätzliche Varianzkomponenten parametrisieren, explizit in das Messmodell mit einbezogen werden können. In der Praxis ist man jedoch daran interessiert, die vergleichsweise einfachen 1PL-, 2PL- oder 3PL-Modelle beizubehalten, die keine Modellierung von Kontexteffekten erlauben. In Large-Scale Assessments werden also teils bewusst „unterkomplexe" oder „fehlspezifizierte" Modelle eingesetzt, und es wird lediglich versucht, die *Auswirkungen* dieser Fehlspezifizierung zu minimieren. Hierbei kommen Testdesigns zum Einsatz.

Praktisch treten dabei zwei grundsätzliche Probleme auf: Das Beispiel der *NAEP reading anomaly* hat gezeigt, dass selbst ein sich gravierend auswirkender Kontexteffekt in IRT-Modellen nicht immer direkt sichtbar ist. Es gibt kein „absolutes" Gütekriterium, das Auskunft darüber gibt, ob das verwendete IRT-Modell substanziell fehlspezifiziert ist (und falls ja, ob daraus praktisch bedeutsame Konsequenzen resultieren) oder ob das verwendete Testdesign inadäquat ist. Die oben genannten Dimensionalitätsprüfungen liefern günstigenfalls ein Indiz über eine nicht gegebene eindimensionale Struktur, jedoch keine Informationen über deren Ursache. Im Falle der *NAEP reading anomaly* konnte das Auftreten eines Kontexteffekts aufgrund schwer zu interpretierender Ergebnisse vorerst lediglich vermutet werden. Erst als der vermutete Kontexteffekt in der 1988er Erhebung in Form einer konkreten Hypothese in einem experimentellen Design explizit modelliert (und quasi „repliziert") wurde, konnte die Inadäquatheit des ursprünglichen Modells beziehungsweise des ursprünglichen Testdesigns sichtbar gemacht werden. Tatsächlich ist es im Rahmen der IRT schwer, eine Modellmisspezifizierung zu erkennen, sofern kein explizit formuliertes alternatives Modell vorliegt, dessen Passung gegen das Referenzmodell verglichen werden kann (Frey & Bernhardt, 2012). Die üblicherweise verwendeten Gütekriterien Infit/Outfit (Adams, Wilson & Wang, 1997; Adams & Wu, 2007; Wu, Adams, Wilson & Haldane, 2007), Prüfung auf differentielles Itemfunktionieren (*differential item functioning, DIF*) (Holland & Wainer, 1993; Penfield & Camilli, 2007), sowie die Prüfung auf lokale stochastische Unabhängigkeit (Jiao, Wang & He, 2013; Monseur et al., 2011; Schroeders, Robitzsch & Schipolowski, 2014; Yen, 1984, 1993) erlauben zumeist nur, einzelne Items oder Itempaare, die dem unterstellten Messmodell nicht entsprechen, zu identifizieren, nicht aber gegebenenfalls das Modell als Ganzes zu evaluieren. Verschiedene Publikationen (Fox & Glas, 2001; Sinharay, 2005; Sinharay & Haberman, 2014; Sinharay & Johnson, 2003; Swaminathan et al., 2007; Yen, 1980) widmen sich daher dem Problem, wie sich praktische Konsequenzen eines potentiell misspezifizierten Modells oder

eines potentiell ungeeigneten Testdesigns quasi explorativ, also ohne Formulierung eines alternativen Modells erkennen lassen. Auf diese Methoden, die aus bayesianischen Verfahren abgeleitet sind, wird im Kapitel 4.2 näher eingegangen.

Der folgende Abschnitt beschreibt Testdesigns zunächst allgemein und geht anschließend auf ihre besondere Bedeutung bei der Behandlung von Kontexteffekten ein.

2.6 Testdesign

Testdesigns sind in Studien notwendig, in denen latente Merkmale über große Mengen manifester Indikatoren (Items) gemessen werden. Solche Studien sind z. B. Large-Scale Assessments, bei denen die zu messenden latenten Konstrukte durch sehr viele manifeste Testitems beschrieben werden, die eine Schülerin oder ein Schüler allein nicht bearbeiten könnte. Jede Schülerin und jeder Schüler bearbeitet demzufolge nur eine Teilmenge aller Items: eine Stichprobe von Personen bearbeitet eine Stichprobe von Items. In der Literatur ist dann häufig von *item sampling* die Rede (Van der Linden, Veldkamp & Carlson, 2004; von Davier, M., Sinharay, Oranje & Beaton, 2007). Das Testdesign spezifiziert diesen Samplingprozess. Üblicherweise wird die Gesamtmenge der Items auf verschiedene Testheftversionen (*test forms*) verteilt. Die Gesamtstichprobe wird dann in mehrere Teilstichproben aufgeteilt, wobei jede Teilstichprobe eine Testheftversion bearbeitet. Man spricht dabei von einem Multi-Matrix-Design (*multiple matrix sampling design*), da für alle Schüler, die eine Testheftversion bearbeitet haben, eine vollständige Datenmatrix resultiert (Frey, Hartig & Rupp, 2009; Gonzalez & Rutkowski, 2010). Die gesamte Datenmatrix ist aus diesen einzelnen, testheftspezifischen Matrizen zusammengesetzt und lückenhaft.

Das durch das Testdesign definierte *item sampling* ist also ein Mittel, mit solchen großen Itemmengen umzugehen. Dabei gibt es nicht nur ein, sondern viele verschiedene Testdesigns, die prinzipiell infrage kommen können. Das Testdesign determiniert die Struktur der Daten, genauer: die Struktur *aller* einzelnen Faktoren (Personen, Items, Testhefte, Itempositionen etc.) zueinander. So kann etwa eine gewünschte Struktur zweier Faktoren zueinander über ein entsprechend gewähltes Testdesign realisiert werden; wenn etwa Items in Testheften genestet sein sollen.

Die wichtigsten Datenstrukturen, die für diese Arbeit Relevanz haben, sollen nun kurz anhand zweier Faktoren (F1 und F2) beschrieben werden. Die jeweils gewählten Beispiele orientieren sich dabei an realen Anwendungsfällen. Zudem wird die Datenstruktur durch eine Kreuztabelle dargestellt, wobei der erste Faktor F1 in den Spalten, der zweite Faktor F2 in den

Zeilen dargestellt ist. Die Werte in der Tabelle bezeichnen die Anzahl der jeweiligen Beobachtungen für diese Kombination der beiden Faktoren.

1. F1 in F2 genestet: Ein einfaches Beispiel dafür sind Schüler (F1) und Klassen (F2). Jede Schülerin oder jeder Schüler gehört zu genau einer Klasse, wobei jede Klasse mehrere Schülerinnen und Schüler enthält. Dargestellt in einer Kreuztabelle, gibt es für zwei genestete Faktoren in jeder Spalte der Tabelle genau ein Element > 0.

F2 \ F1	S1	S2	S3	S4	S5	S6	S7	S8	S9
A	1	1	1						
B				1	1	1			
C							1	1	1

   In diesem Beispiel gibt es also drei Klassen A, B und C mit jeweils 3 Schülerinnen oder Schülern.

2. F1 und F2 partiell gekreuzt: Ein einfaches Beispiel dafür sind Schüler (F1) und Items (F2). Jede Person bearbeitet eine Teilmenge aller Items, wobei sich die Teilmengen, allerdings nicht vollständig, überschneiden. Eine solche Datenstruktur zwischen Personen und Items resultiert üblicherweise aus Multi-Matrix-Designs. Formal steht in jeder Spalte der Matrix mindestens ein Element > 0 und in mindestens einer Spalte stehen mindestens 2 Elemente > 0.

F2 \ F1	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
I1	1	1	1				1	1	1			
I2	1	1	1	1	1	1						
I3	1	1	1							1	1	1
I4										1	1	1
I5				1	1	1	1	1	1	1	1	1

   In diesem Beispiel gibt es 5 Items (I1, I2, I3, I4 und I5) und 12 Schüler. Schüler S01, S02 und S03 bearbeiten die Items I1, I2 und I3, Schüler S04, S05 und S06 die Items I2 und I5, usw.

3. F1 und F2 vollständig gekreuzt: Ein einfaches Beispiel dafür sind die beiden Faktoren Klassen (F1) und Testhefte (F2). Für jede Kombination der Faktorstufen beider Faktoren gibt es mindestens eine Beobachtung.

F2 \ F1	K1	K2	K3	K4	K5	K6
TH1	5	3	5	4	4	6
TH2	5	5	5	5	4	6
TH3	4	4	5	4	4	6
TH4	5	4	4	5	4	7

In diesem Beispiel gibt es 4 Testhefte (TH1, TH2, TH3, TH4) und 6 Klassen. Jedes der Testhefte wird in jeder Klasse mehrmals eingesetzt, wobei beispielsweise Testheft TH1 in Klasse K1 fünfmal eingesetzt wird.

4. F1 und F2 balanciert: Für jede Kombination der Faktorstufen beider Faktoren gibt es dieselbe Anzahl an Beobachtungen. Ein Beispiel dafür sind z. B. Items (F1) und Itempositionen (F2). Jedes Item tritt an jeder Position mit derselben Häufigkeit auf.

F2 \ F1	**I01**	**I02**	**I03**	**I04**	**I05**	**I06**	**I07**	**I08**	**I09**	**I10**	**I11**	**I12**
P1	200	200	200	200	200	200	200	200	200	200	200	200
P2	200	200	200	200	200	200	200	200	200	200	200	200
P3	200	200	200	200	200	200	200	200	200	200	200	200
P4	200	200	200	200	200	200	200	200	200	200	200	200

In diesem Beispiel gibt es 12 Items und 4 Itempositionen (P1, P2, P3, P4). Jedes Item kommt an jeder Itemposition genau 200-mal vor.

Balancierung bedeutet damit, dass die auf F1 bedingte Verteilung von F2 – also F2 | F1 – für alle Faktorstufen von F1 gleich ist. (Die Verteilung von Positionen ist für jedes Item gleich.) Ebenso ist die auf F2 bedingte Verteilung von F1 – also F1 | F2 – für alle Faktorstufen von F2 gleich. (Die Verteilung von Items ist für jede Position gleich.) F1 und F2 sind damit orthogonal zueinander: $F1 \perp F2$. Das bedeutet zugleich, dass die Kovarianz von F1 und F2 Null ist: $cov(F1, F2) = 0$. Aus der Balancierung folgt Orthogonalität.

5. F1 und F2 miteinander verlinkt: Hierbei handelt es sich nicht um eine Datenstruktur, sondern eine mögliche Eigenschaft partiell gekreuzter Faktoren (siehe Datenstruktur Nr. 2). Zwei partiell gekreuzte Faktoren können, müssen aber nicht miteinander verlinkt sein. Zwei Faktoren (z. B. Personen und Items) über einen weiteren Faktor (z. B. Testhefte) miteinander zu verlinken, ist ein gängiges Vorgehen bei der Konstruktion von Testdesigns (von Davier, A. et al., 2008) und soll hier kurz erwähnt werden. Verlinken bedeutet, dass sämtliche Items direkt oder indirekt miteinander verbunden sind. Zwei Items A und B sind direkt miteinander verbunden, wenn beide Items von derselben Person bearbeitet werden, also im selben Testheft auftreten. Zwei Items A und B sind indirekt (z. B. über ein weiteres Item C) miteinander verbunden, wenn beide Items zwar nicht in demselben Testheft auftreten, es aber zwei Testhefte gibt, wobei in einem Testheft Item A gemeinsam mit Item C, und in dem zweiten Testheft Item B gemeinsam mit Item C auftritt. Bei dem unter 2.) dargestellten Beispiel zweier partiell gekreuzter Faktoren handelt es sich zugleich um ein Design, in dem beiden Faktoren F1 und F2 miteinander verlinkt sind.

Die hier beschriebenen Datenstrukturen beschreiben immer die Beziehung zweier Faktoren zueinander und sind hierarchisch geordnet. So sind zwei genestete Faktoren niemals vollständig miteinander verlinkt, zwei vollständig gekreuzte Faktoren sind es immer. Faktoren in einer partiell gekreuzten Datenstruktur können, aber müssen nicht vollständig miteinander verlinkt sein (vgl. auch Hecht, Weirich, Siegle & Frey, 2015).

Die Abhängigkeit der oben beschriebenen Datenstrukturen bedeutet etwa, dass verlinkte Designs immer (mindestens partiell) gekreuzt, und balancierte Designs immer vollständig gekreuzt sind.

Gibt es mehr als zwei Faktoren in den Daten (z. B. Items, Personen, Testhefte und Klassen), kann für jede paarweise Kombination aus diesen Faktoren eine andere Struktur bestehen: So können Personen in Klassen genestet sowie Personen und Items partiell gekreuzt sein, woraus folgt, dass Klassen und Items entweder partiell oder vollständig gekreuzt sein müssen. Ein und dasselbe Testdesign kann also bezüglich Personen und Items partiell gekreuzt, bezüglich Personen und Klassen genestet und bezüglich Items und Itempositionen balanciert sein.

2.7 Bedeutung des Testdesigns: Itempositionseffeke

Am Ende des Kapitels 2.1 stand die Frage, welche Möglichkeiten existieren, aus dem Raschmodell unverzerrte Parameter zu gewinnen, obwohl es in den meisten Anwendungsfällen im strengen Sinne ein fehlspezifiziertes Modell ist. Zentrale Bedeutung hat hierbei das Testdesign, was im Folgenden näher ausgeführt werden soll. Es wird dabei auf Itempositionseffekte als ein Beispiel für Kontexteffekte zurückgegriffen, obwohl die erläuterte Balancierungsmethode prinzipiell auf verschiedene Kontexteffekte anwendbar ist.

Angenommen, die Wahrscheinlichkeit $P(X_{nip}=1)$, ein Item korrekt zu lösen, sei durch genau drei Faktoren (oder Varianzkomponenten) systematisch bestimmt: (a) der Schwierigkeit β_i des Items (je größer die Schwierigkeit β_i, desto geringer $P(X_{nip}=1)$), (b) der Fähigkeit θ_n der Person, die das Item zu lösen versucht (je größer die Fähigkeit θ_n, desto höher $P(X_{nip}=1)$), sowie (c) der Position λ_p des Items im Testheft (je weiter hinten das Item im Testheft steht, also je „größer“ sein Positionsindex ist, desto geringer ist $P(X_{nip}=1)$). Der Index $n=\{1,\dots,N\}$ bezeichnet dabei die jeweilige Person, der Index $i=\{1,\dots,I\}$ das jeweilige Item, und der Index $p=\{1,\dots,P\}$ die jeweilige Position des Items im Testheft. Wenn es etwa in einem Testdesign 1500 Personen, 120 Items und vier mögliche Positionen gibt, an denen ein Item im Testheft auftreten kann, so kann n Werte von 1 bis 1500, i Werte von 1 bis 150, und p Werte von 1 bis 4 annehmen. Entsprechend wäre $N=1500$, $I=120$ und $P=4$, und es

können $P-1=3$ Positionseffekte parametrisiert werden: der Effekt der zweiten, dritten und vierten Position, wobei jeder Effekt in Relation zur ersten Position ausgedrückt werden würde. Das entsprechende „wahre“ Modell würde dann so aussehen:

$$\text{logit}\left(P(X_{nip}=1)\right)=\theta_n-\beta_i+\lambda_p. \tag{4}$$

Das Raschmodell aus Gleichung 1 (Seite 13) nimmt an, dass der Effekt von λ_p gleich Null ist, also $\lambda_p=0$. Mit anderen Worten: egal, an welcher Position ein Item im Testheft platziert wird, die Schwierigkeit des Items ist an allen Positionen gleich. Im Raschmodell wird nicht modelliert und dementsprechend auch nicht getestet, ob dieser Effekt tatsächlich nicht auftritt; $\lambda_p=0$ wird stattdessen vorausgesetzt. Das Raschmodell wäre hier demzufolge (wie in den meisten Anwendungsfällen) ein unterkomplexes, misspezifiziertes Modell; es modelliert die Wahrscheinlichkeit $P(X_{nip}=1)$ nicht hinreichend. In verschiedenen empirischen Anwendungsfällen ist man jedoch an einer vollständigen Modellierung von $P(X_{nip}=1)$ gar nicht interessiert; es gilt lediglich, den Effekt von θ_n und β_i auf diese Wahrscheinlichkeit unverfälscht schätzen zu können. Die zentrale Frage ist daher: Angenommen, Gleichung 4 mit $\lambda_p\neq 0$ sei für einen bestimmten Anwendungsfall das „wahre“ Modell, für die Modellierung würde jedoch das Raschmodell zugrunde gelegt. Wie kann sichergestellt werden, dass der Parameter θ_n aus Gleichung 1 dem „wahren“ Parameter θ_n aus Gleichung 4 entspricht, und dass der Parameter β_i aus Gleichung 1 dem „wahren“ Parameter β_i aus Gleichung 4 entspricht? Oder in anderen Worten: Welche Bedingungen müssen erfüllt sein, um aus einem ggf. invaliden Modell valide Parameterschätzer zu gewinnen? Dazu gibt es mehrere Möglichkeiten (für einen Überblick, siehe Yousfi & Böhme, 2012), von denen hier zwei vorgestellt werden sollen, die in Large-Scale Assessments am häufigsten realisiert werden:

1. *Standardisierung:* Ziel ist es, die Varianz von λ_p zu minimieren. Denkt man sich λ_p als einen Kontexteffekt (wie etwa die Tageszeit der Testung, die Testadministration, die Länge der Testung) mit $\lambda_p\neq 0$, so bedeutet Standardisierung, dass alle Testpersonen denselben Wert auf der Kontextvariable haben; alle Personen bearbeiten den Test zur selben Tageszeit. Selbst wenn es einen signifikanten Effekt der Tageszeit auf die Lösungswahrscheinlichkeit der Items gäbe, wäre dieser Effekt für sämtliche Testteilnehmer konstant und würde zu keiner Verfälschung der Item- und Personenparameter führen. Im Sinne von Brennan (1992) hieße Standardisierung, die „universes of generalization“ so zu reduzieren, dass sie jeweils nur noch eine Bedingung enthalten: nur eine Tageszeit für alle Personen, nur eine Reihenfolge für alle Items, nur eine Position für jedes Item im Test, etc.

2. *Ausbalancieren:* Bestimmte Kontextbedingungen können im Rahmen von Large-Scale Assessments nicht konstant gehalten werden, etwa die Position eines Items im Testheft. Da mehrere Testhefte mit einer unterschiedlichen Auswahl von Items eingesetzt werden müssen, variiert die Reihenfolge von Items notwendigerweise zwischen Testheften. Die Itemposition lässt sich nicht konstant halten, wenn zugleich eine Verlinkung von Items beabsichtigt wird. Sollte also die Position eines Items im Testheft ebenfalls als ein Kontexteffekt die Lösungswahrscheinlichkeit des Items beeinflussen, werden die Itemparameterschätzer immer durch diesen Effekt beeinflusst sein. Die Parameter werden also immer auf irgendeine Weise „verfälscht" sein. Über das Testdesign kann nun lediglich gewährleistet werden, dass diese Verfälschung *sämtliche Items in derselben Weise* beeinflusst. Wenn Items und Itempositionen zueinander balanciert sind (Bedingung 4 in Kapitel 2.6), tritt jedes Item an jeder Position des Testhefts mit gleicher Häufigkeit auf. Selbst wenn die Varianz von λ_p größer als Null ist, es also einen bedeutsamen Effekt der Itemposition gibt, wird sichergestellt, dass dieser Effekt *im Mittel* alle Items in derselben Weise beeinflusst und also keine *spezifische* Verfälschung der Itemparameter auftritt. In einem strengen Sinne trifft dies jedoch nur zu, wenn der Effekt der Itemposition nicht von den Items selbst (oder Eigenschaften der Items) abhängt. Balancierung gewährleistet Orthogonalität und damit, dass die interessierenden Effekte der Personen und der Items *unabhängig* von den „störenden" Kontexteffekten sind.

Die Methode des Ausbalancierens ist eng angelehnt an Methoden aus der Kausalitätsforschung (Campbell & Stanley, 1963; Holland, 1986; Rubin, 1974) und kann analog dazu betrachtet werden. Um Effekte eines Treatments X (z. B. der Gabe eines Medikaments oder eines Placebos) unabhängig von weiteren „störenden" Kovariaten interpretieren zu können, müssen die Treatmenteffekte unabhängig oder orthogonal zu den Effekten der Kovariaten sein. Nur wird Orthogonalität im Rahmen klassischer psychologischer Experimente zumeist nicht durch Ausbalancieren, sondern durch Randomisierung erreicht. Ein Beispiel soll das verdeutlichen:

Der Effekt eines dichotomen Treatments X (1 = Gabe eines Medikaments; 0 = Gabe eines Placebos) auf das Wohlbefinden Y soll in einer einfachen Regression geschätzt werden. Angenommen sei ferner, „in Wahrheit" hänge das Wohlbefinden noch von einer weiteren Kovariaten, dem Geschlecht Z ab, die jedoch in dem Regressionsmodell nicht berücksichtigt wird (man hätte es, genau wie beim Raschmodell, mit einem unterkomplexen, fehlspezifizierten Modell zu tun). Angenommen sei, das „wahre" Modell laute $Y = \alpha_0^w + \alpha_1^w X + \alpha_2^w Z + \varepsilon$ (mit

$\alpha_2^w \neq 0$), das geschätzte Regressionsmodell jedoch $Y = \alpha_0^g + \alpha_1^g X + \varepsilon$. Man kann hier ganz analog die Frage stellen: Unter welchen Bedingungen entspricht das geschätzte α_1^g aus dem Regressionsmodell dem wahren α_1^w? Beispielsweise wäre dies im randomisierten Zufallsexperiment der Fall. Bei zufälliger Treatmentzuweisung sind *X* und *Z* orthogonal zueinander (beziehungsweise unabhängig voneinander). Ferner ist die Kovarianz zwischen *X* und *Z* gleich Null: $cov(X, Z) = 0$. Selbst wenn *Z* einen Effekt auf *Y* hat, muss dieser Effekt im Regressionsmodell nicht mit spezifiziert werden, um einen unverfälschten Schätzer des Effekts für *X* zu gewinnen. Etwas Ähnliches wünscht man sich für das Raschmodell: Selbst wenn es einen Effekt von λ_p auf $P(X_{nip} = 1)$ gibt, muss dieser Effekt im Raschmodell nicht mit spezifiziert werden, um unverfälschte Schätzer der Effekte für θ_n und β_i zu gewinnen. Genau wie die Randomisierung im Zufallsexperiment wird über die Ausbalancierung in Testdesigns Orthogonalität zwischen den Effekten, die im Modell berücksichtigt werden, und den Effekten, die im Modell nicht berücksichtigt werden, angestrebt. Die Ausbalancierung bewirkt – genau wie die Randomisierung –, dass Items und Itempositionen zueinander unkorreliert sind. Dabei ist ein wichtiger Unterschied zu nennen: Im Zufallsexperiment gewährleistet die Randomisierung nicht nur, dass *X* und *Z* unabhängig voneinander sind, sondern dass *X* außerdem zu *jeder weiteren denkbaren Kovariate unabhängig ist.* Das ist im Raschmodell anders. Hier wird Unabhängigkeit nicht durch Randomisierung realisiert[1], sondern durch systematische Gleichverteilung der Items auf alle Positionen. Damit sind lediglich Items und Positionen unabhängig zueinander, nicht aber Items und weitere mögliche Kontextvariablen, wie etwa Testhefte.

Da Orthogonalität im Raschmodell (innerhalb von Large-Scale Assessments) nicht über Randomisierung realisiert werden kann, bedeutet dies, man müsste vorab jeden möglichen Kontexteffekt berücksichtigen und paarweise zu den Items ausbalancieren. Praktisch ist das jedoch nicht möglich, da die Menge an dazu benötigten Testheften nicht realisiert werden kann.

Da keine Möglichkeit existiert, in einem Testdesign alle potentiellen Kontexteffekte gleichzeitig auszubalancieren, müssen vorab diejenigen Kontextvariablen identifiziert werden, die in praktischen Anwendungen substanzielle Effekte auf $P(X_{nip} = 1)$ ausüben und gegebenenfalls die Effekte von θ_n und β_i verfälschen können. Für diese Kontextvariablen müssen dann adäquate Balancierungsmethoden gefunden und empirisch überprüft werden.

Das bis hierhin Gesagte soll kurz zusammengefasst werden:

[1] Randomisierung würde ja bedeuten, dass jedes Testheft eine zufällige Auswahl der Items in einer zufälligen Reihenfolge enthalten müsste. Diese Praxis ist nicht umsetzbar, da einerseits jede Person ein individuelles Test-

- Einfache IRT-Modelle (wie das Raschmodell) sind häufig unterkomplex, berücksichtigen also nicht alle Kontextvariablen, die die Wahrscheinlichkeit $P(X_{nip}=1)$ beeinflussen. Aus den Modellgüteindizes kann nicht abgelesen werden, ob ein solcher Kontexteffekt auftritt oder deswegen praktische Konsequenzen für die Güte der interessierenden Parameter zu befürchten sind. Um Kontexteffekte erkennen zu können, müssen sie explizit modelliert werden.
- Um aus unterkomplexen, Kontexteffekte nicht berücksichtigenden Modellen dennoch valide Parameter zu gewinnen, können Methoden der *Standardisierung* und *Ausbalancierung* benutzt werden. Dazu müssen die Kontexteffekte, für die das Testdesign ausbalanciert werden soll, jedoch bekannt sein: Eine Randomisierung wie im Zufallsexperiment, die für alle denkbaren Kontexteffekte Balancierung gewährleistet, ist nicht möglich.
- Da ferner nicht alle potentiell auftretenden Kontexteffekte ausbalanciert werden können, ist es wichtig zu prüfen, welche Kontexteffekte einen substanziellen Effekt auf $P(X_{nip}=1)$ ausüben, um anschließend Testdesigns entwickeln und einsetzen zu können, die eine Ausbalancierung für eben diese Kontexteffekte erlauben.
- Ein erster Schritt ist daher die Entwicklung von Modellen, die eine zuverlässige Messung solcher Kontexteffekt erlauben.

Dem letzten hier genannten Punkt widmet sich der erste Einzelbeitrag der Dissertation. Als ein Beispiel für Kontexteffekte wird dabei die Modellierung von Itempositionseffekten behandelt.

3. Zusammenfassung des ersten Einzelbeitrags: Modeling Item Position Effects Using Generalized Linear Mixed Models

Dieser Beitrag ist in der Zeitschrift *Applied Psychological Measurement* erschienen. Die Referenz lautet:

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*(7), 535-548. doi: 10.1177/0146621614534955

Der erste Einzelbeitrag behandelt die Modellierung (unidimensionaler) Itempositionseffekte mittels eines linear-logistischen Testmodells (LLTM). Die Besonderheit dieses Modells ist,

---

heft erhalten müsste, und zum zweiten die Testitems häufig an Aufgabenstimuli gebunden sind und daher in ih-

dass es – abgeleitet von allgemeinen linearen gemischten Modellen (*generalized linear mixed models*, GLMM) (De Boeck & Wilson, 2004; Wilson & De Boeck, 2004) – die Spezifizierung von Fehlertermen erlaubt. In Anlehnung an De Boeck et al. (2011) wurde das Modell zur Abgrenzung gegen das originale LLTM von Fischer (1973) als LLTM + ε bezeichnet. Es wurde erwartet, dass die von De Boeck et al. (2011) theoretisch begründeten und von Hohensinn et al. (2008) empirisch vorgefundenen Probleme des originalen LLTM bezüglich der Verzerrtheit des nominellen Alphaniveaus für das LLTM + ε nicht auftreten. Darüber hinaus sollte geprüft werden, ob dieser Befund in gleicher Weise für vollständig balancierte, teilweise balancierte und unbalancierte Testdesigns auftritt. Zusätzlich wurde untersucht, ob das Raschmodell zumindest in vollständig balancierten Designs unverfälschte Parameter schätzt, selbst wenn aufgrund des Auftretens von Itempositionseffekten das Raschmodell formal ein fehlspezifiziertes Messmodell darstellt.

In einer Simulationsstudie konnte gezeigt werden, dass das LLTM + ε das nominelle Alphaniveau von 5% einhält und eine hinreichend gute Testpower gewährleistet. Beides ist jedoch nur in Testdesigns der Fall, die bezüglich der Itemposition vollständig balanciert waren. Für eine unverzerrte Schätzung kommt es also nicht nur auf die Adäquatheit des Messmodells, sondern *zusätzlich* auf die Adäquatheit des gewählten Testdesigns an.

Ferner wurde untersucht, ob die Parameter im Raschmodell durch das Auftreten von Itempositionseffekten verzerrt werden. Erwartungsgemäß zeigte sich, dass der Bias in den Itemparametern in teilweise balancierten Designs geringer war als in unbalancierten Designs, und in vollständig balancierten Designs geringer als in teilweise balancierten Designs. Entgegen der Erwartung war der Bias selbst in vollständig balancierten Designs größer als Null und nahm mit ansteigendem Positionseffekt ebenfalls zu. Die Ursache dieses Phänomens wird in Kapitel 4 näher erläutert.

4. Zum Problem der Varianzeinschränkung in unterkomplexen IRT-Modellen

In dem folgenden Kapitel soll auf ein Problem eingegangen werden, das in dem ersten Einzelbeitrag zwar thematisiert, jedoch nicht erschöpfend diskutiert worden ist: In der Simulationsstudie wurden Daten beruhend auf einem Modell mit Positionseffekten (entsprechend Gleichung 4, siehe Seite 28) simuliert und anschließend beruhend auf einem unterkomplexen

rer Auswahl nicht zufällig zwischen Testheften variieren können.

Modell (entsprechend Gleichung 1, siehe Seite 13) analysiert. Das Messmodell „unterschlägt" damit die Varianz aufgrund von Positionseffekten, es berücksichtigt nicht alle Varianzkomponenten (vgl. Kapitel 2.1). Die Annahme war: In einem Design, in dem die beiden Faktoren Items und Positionen balanciert sind (*balanced incomplete block design*; BIB, siehe Bedingung 4 auf Seite 26), ist der im Modell 1 nicht berücksichtigte Effekt der Itemposition λ_p über die im Modell berücksichtigten Effekte von β_i und θ_n ausbalanciert, die beiden Faktoren Items und Itempositionen sind orthogonal zueinander (vgl. Kapitel 2.6 und 2.7). Der durch Modell 1 geschätzte Effekt von β_i sollte unverzerrt sein, also keinen Bias zeigen. In Tabelle 3 auf Seite 544 des ersten Einzelbeitrags zeigt sich jedoch, dass dies nicht zutrifft: Es tritt ein zwar sehr kleiner, jedoch mit steigendem Positionseffekt ebenfalls ansteigender Bias auf. Die Ursache dieses Bias liegt darin begründet, dass in linearen und log-linearen Regressionsmodellen die Metrik der Skalen, auf die sich die Parameter beziehen, unterschiedlich ist. Dies soll anhand desselben Beispiels wie in Kapitel 2.7 näher erläutert werden.

Das Wohlbefinden Y soll in einer einfachen Regression aus dem Treatment X (z. B. der Gabe eines Medikaments ($X=1$) oder eines Placebos ($X=0$)) vorhergesagt werden. Angenommen sei ferner, „in Wahrheit" hänge das Wohlbefinden zusätzlich noch von einer weiteren Kovariaten, dem Geschlecht Z ab. Das Wohlbefinden sei dabei einfach auf einer siebenstufigen Likert-Skala gemessen, deren Spannweite von 1 = sehr schlecht bis 7 = sehr gut reicht. Das „wahre" Modell laute $Y = \alpha_0^w + \alpha_1^w X + \alpha_2^w Z + \varepsilon^w$ (mit $\alpha_2^w \neq 0$), das geschätzte Regressionsmodell jedoch $Y = \alpha_0^g + \alpha_1^g X + \varepsilon^g$. Wenn $X \perp Z$ und $\alpha_2^w \neq 0$, dann gilt für die Varianz der Fehlerterme ε^w und ε^g auch immer: $var(\varepsilon^w) < var(\varepsilon^g)$. In einem randomisierten Zufallsexperiment wären X und Z ausbalanciert, folglich $\alpha_1^g = \alpha_1^w$. Angenommen man findet einen unstandardisierten Effekt von $\alpha_1^g = 0.5$. Für die Interpretation des Effekts bedeutet dies: Wenn eine Behandlung mit dem Medikament stattfindet ($X=1$), ist das erwartete Wohlbefinden (auf der siebenstufigen Likert-Skala) um 0.5 Punkte höher, als wenn eine Behandlung mit einem Placebo stattfindet. Die Metrik des Effekts von α_1^g hängt also vom Skalenniveau von Y ab; der Effekt wird relativ zur Varianz des Kriteriums Y ausgedrückt. Diese Kriteriumsvarianz bleibt dabei immer gleich, egal ob nun weitere Kovariaten (z. B. Z) in das Modell mit aufgenommen werden oder nicht. Die Metrik von α_1^g und α_1^w ist also dieselbe, und die Interpretation des Effekts von α_1^g hängt nicht davon ab, ob und falls ja, wie viele weitere Kovariaten durch das Modell parametrisiert werden.

Nun soll derselbe Fall für die beiden logistischen Regressionsmodelle betrachtet werden; das durch Gleichung 4 beschriebene „wahre“ Modell: $\text{logit}\big(P(X_{nip}=1)\big)=\theta_n-\beta_i+\lambda_p$, sowie das durch Gleichung 1 beschriebene und für die Parameterschätzung benutzte Raschmodell: $\text{logit}\big(P(X_{ni}=1)\big)=\theta_n-\beta_i$. In logistischen Regressionsmodellen werden die Parameter nicht in Relation zur Varianz des Kriteriums, sondern relativ zur standard-logistischen Varianz ausgedrückt. Diese nimmt für Modelle, die die logistische Transformationsfunktion (siehe Gleichung 2, Seite 13) verwenden, den Wert 3.29 an. Während in linearen Regressionsmodellen die Varianz des Kriteriums *Y* immer gleich bleibt und die Fehlervarianz variiert (also geringer wird, wenn weitere Prädiktoren in das Modell aufgenommen werden), variiert in logistischen Regressionsmodellen die Varianz des Kriteriums zwischen Modellen, während die standard-logistische Varianz (die „Fehlervarianz“) gleich bleibt. Das bedeutet: In klassischen linearen Regressionsmodellen verändert sich die Metrik eines Prädiktors (und damit die Interpretation des Effekts) nicht, wenn weitere Prädiktoren in das Modell aufgenommen werden. In IRT-Modellen jedoch verändert sich die Metrik und damit auch die Interpretation eines Effekts (z. B. β_i), wenn zusätzlich weitere Effekte (z. B. λ_p) parametrisiert werden. Das ist immer dann der Fall, wenn diese zusätzlichen Effekte (z. B. λ_p) eine Varianz größer als Null haben. Praktisch bedeutet das: Auch in vollständig ausbalancierten Designs muss mit einem Bias gerechnet werden, sofern das Messmodell nicht alle relevanten Varianzkomponenten berücksichtigt.

Dieses Problem wurde bislang nur wenig und vorwiegend in soziologischer Literatur zur logistischen Regression behandelt (Allison, 1999; Mood, 2010; Woods & Harpole, 2014). Mood (2010) diskutiert dieses Phänomen ausführlich, wobei sie von „unobserved heterogenity“ anstatt von „nicht berücksichtigten Varianzkomponenten“ spricht. Sie führt aus, dass „unobserved heterogenity“ in logistischen Regressionsmodellen zu irreführenden Interpretationen führen kann, wenn etwa Parameter zweier verschiedener Personengruppen oder Parameter zweier verschiedener Modelle für dieselben Daten oder Parameter desselben Modells für zwei verschiedene Datensätze verglichen werden sollen. Der zentrale Punkt dabei ist, dass in der logistischen Regression nicht berücksichtigte Komponenten auch dann verzerrte Parameter erzeugen, wenn sie nicht mit den interessierenden Effekten zusammenhängen. Dasselbe Prinzip kann auf die IRT und Large-Scale Assessments übertragen werden: Kontexteffekte können selbst bei ausbalanciertem Design zu verzerrten Parametern im Raschmodell führen, obwohl die Unabhängigkeit von interessierenden Effekten (der Items und der Personen) und Kontexteffekten durch Balancierung gegeben ist. In den folgenden zwei Kapiteln 4.1 und 4.2 soll daher darauf eingegangen werden, in welcher Weise diese Verzerrungen in IRT-Modellen

auftreten können, und inwiefern einer der von Mood (2010) vorgeschlagenen Lösungsansätze auch im IRT-Kontext anwendbar sind. Zwei konkrete Fragen sollen dabei im Vordergrund stehen:

1. Wie gravierend ist der Bias aufgrund der Varianzeinschränkung?
2. Wie lässt sich in einem empirischen Fall erkennen, ob in dem spezifizierten Messmodell relevante Varianzkomponenten nicht mit modelliert werden und also die Gefahr verzerrter Parameterschätzer besteht?

4.1 Bias aufgrund der Varianzeinschränkung

Im vorangegangenen Abschnitt wurde beschrieben, dass die Gesamtvarianz in linearen Regressionsmodellen stets bekannt ist, während die Größe der Fehlervarianz von dem jeweiligen Modell abhängt. In der IRT hingegen ist die Größe der „Fehlervarianz" stets gleich, während die Größe der Gesamtvarianz von dem jeweiligen Modell abhängt. Das bedeutet: Um bestimmen zu können, ob und falls ja, wie stark ein Bias im Messmodell auftritt, muss das „wahre" (oder „vollständige") Modell bekannt sein. Dies soll wiederum an dem Beispiel der beiden Modelle aus Gleichung 1 und Gleichung 4 veranschaulicht werden:

Angenommen, die empirischen Daten beruhen auf dem Modell entsprechend Gleichung 4: $\text{logit}\big(P(X_{nip}=1)\big)=\theta_n-\beta_i+\lambda_p$, wobei willkürlich für alle Parameter eine „wahre" Varianz von 1 angenommen werden soll: $var^w(\theta_n)=1$, $var^w(\beta_i)=1$ und $var^w(\lambda_p)=1$. Diese Daten werden nun mit dem Raschmodell $\text{logit}\big(P(X_{ni}=1)\big)=\theta_n-\beta_i$ geschätzt. Die durch das „wahre" Modell abgebildete Gesamtvarianz var^w ist

$$var^w=var^w(\theta_n)+var^w(\beta_i)+var^w(\lambda_p)+3.29. \quad (5)$$

In diesem Beispiel wäre $var^w=1+1+1+3.29=6.29$. Der relative Varianzanteil jeder Modellkomponente lässt sich durch Multiplikation mit $(var^w)^{-1}$ bestimmen. In dem Modell würden also $\frac{1}{6.29}\approx 15.9\%$ der Varianz in den Itemantworten darauf zurückzuführen sein, dass die Items eine unterschiedliche Schwierigkeit haben. Da $var^w(\lambda_p)=1$, würden ebenfalls 15.9% der Varianz in den Itemantworten darauf zurückzuführen sein, dass jedes einzelne Item an unterschiedlichen Positionen eine unterschiedliche Schwierigkeit (Lösungshäufigkeit) hat. Um zu bestimmen, wie groß der Bias näherungsweise für Items ausfällt, wenn die Daten fälschlicherweise mit dem Raschmodell analysiert werden (wenn also λ_p nicht parametrisiert wird), kann die durch das Raschmodell „fehlspezifizierte" Gesamtvarianz var^g und anschließend der relative Anteil der Itemvarianz an dieser Gesamtvarianz bestimmt werden:

$$var^{g}(\beta_i) = \frac{var^{w}(\beta_i)}{var^{w}} \cdot 3.29 \cdot \left(1 - \frac{var^{w}(\beta_i) + var^{w}(\theta_n)}{var^{w}}\right)^{-1} . \quad (6)$$

Setzt man für $var^{w} = 6.29$ ein, so würde man im Raschmodell für Items eine Varianz von 0.767 schätzen, obwohl die Itemvarianz in Wahrheit 1 ist. Ein Item mit einem „wahren" Logitwert von 1.11 hätte folglich einen Logitwert von 0.97 und einen Bias von 0.14.

Zur Illustration soll dieses Beispiel nun auf empirische Daten des IQB-Ländervergleichs 2011 in der Primarstufe (Stanat et al., 2012) im Fach Deutsch für den Kompetenzbereich Lesen übertragen werden. Es wird dabei auf dieselben Daten zurückgegriffen, die im Anhang des ersten Einzelbeitrags Grundlage des empirischen Beispiels waren. Aus der Gesamtstichprobe des Ländervergleichs 2011 von 26 522 Schülerinnen und Schülern (49.2% weiblich; mittleres Alter: 10.46 Jahre) wurde eine repräsentative (das heißt, an der Größe der Bundesländer gewichtete) und bezüglich der Itemposition vollständig balancierte Teilstichprobe von 4000 Personen gezogen. In diese Stichprobe gingen nur Personen des Hauptdesigns ein (vgl. Richter et al., 2012), das bedeutet, es wurden weder Förderschüler noch Schüler, die Testhefte mit Orthographieaufgaben bearbeitet haben, in die Ziehung einbezogen. Der Test selbst beinhaltete 80 dichotom kodierte Items, die in 11 Aufgaben (bzw. Testlets) genestet waren. Jeweils zwei Aufgaben wurden in einem Testheft zusammengestellt. In jedem Testheft gab es vier mögliche Blockpositionen, an denen eine Aufgabe auftreten kann. Da die Position eines Items *innerhalb* einer Aufgabe unveränderlich war, entsprach die Position eines Items der Blockposition der jeweiligen Aufgabe im Testheft. Zwei Items einer Aufgabe, die an unterschiedlichen Positionen innerhalb der Aufgabe auftreten, haben demzufolge in der hier modellierten Variante denselben Wert auf der Positionsvariablen, nämlich den der Blockposition der Aufgabe im Testheft. Es wird nun für diese Daten zum einen ein Raschmodell und zum anderen ein Modell spezifiziert, das zusätzlich Positionseffekte als Zufallseffekte parametrisiert. Dabei soll angenommen werden, bei dem Modell mit Positionseffekten handle es sich um das wahre Modell.[2] Alle Varianzkomponenten werden dabei als zufällig (*random*) modelliert; für das Raschmodell entspricht dies der random person–random item (RPRI) Formulierung (De Boeck, 2008).

Raschmodell: $\text{logit}\left(P(X_{ni} = 1)\right) = \theta_n - \beta_i$

Modell mit Positionseffekten: $\text{logit}\left(P(X_{nip} = 1)\right) = \theta_n - \beta_i + \lambda_p + \rho_{i \times p}$

[2] Diese Annahme wird hier nur der Anschaulichkeit zuliebe getroffen; sie ist in den empirischen Daten nicht zutreffend.

Das Modell mit Positionseffekten schätzt dabei zwei zusätzliche Parameter, λ_p für den Positionseffekt, und $\rho_{i\times p}$ für die Interaktion aus Positions- und Itemeffekt. Die Ergebnisse beider Modelle sind in Tabelle 1 abgebildet.

Tabelle 1: Varianzkomponenten im Raschmodell und Modell mit Positionseffekten

Parameter	**Raschmodell**			**Modell mit Positionseffekten**		
Fixed effects	Est.	SE	*p*	Est.	SE	*p*
(Intercept)	0.562	0.132	< .001	0.560	0.140	< .001
Random effects	Var	SD		Var	SD	
Personen	1.019	1.009		1.020	1.010	
Items	1.487	1.219		1.500	1.225	
Positionen				0.010	0.100	
Positionen × Items				0.010	0.100	
Model Fit						
AIC	75 008			74 954		
BIC	75 035			74 999		
deviance	75 002			74 944		

Verglichen mit den Effekten auf der Item- und Personenseite sind die Effekte der Itemposition und die Interaktion aus Item- und Positionseffekt sehr gering. Unter der Annahme, das Modell mit Positionseffekten sei das wahre Modell, betrüge die Gesamtvarianz $1.02 + 1.50 + 0.01 + 0.01 + 3.29 = 5.83$. Unter derselben Annahme lässt sich nun nach Gleichung 6 (Seite 36) bestimmen, wie groß die geschätzte Varianz für Items $var^w(\beta_i)$ im fehlspezifizierten Raschmodell ausfallen würde, das weder Effekte der Itemposition noch die Interaktion berücksichtigt:

$$\mathrm{var}^g(\beta_i) = \frac{1.50}{5.83} \cdot 3.29 \cdot \left(1 - \frac{1.50 + 1.02}{5.83}\right)^{-1} = 1.49\,.$$

Dies entspricht tatsächlich auch der hier geschätzten Itemvarianz im Raschmodell. Die Varianzeinschränkung beträgt demzufolge $1.50 - 1.49 = 0.01$. Ein Item mit einem „wahren“ Logitwert von 1.11 hätte folglich einen Logitwert von 1.1066 und einen Bias von 0.0034. Dieser Bias ist extrem gering, da die Varianz der Itempositionseffekte (in Relation zur Varianz der Items selbst) ebenfalls äußerst gering ist. Der Bias wäre aber nur dann tatsächlich so gering,

wenn zwei Bedingungen erfüllt sind: (a) bei $\text{logit}\big(P(X_{nip}=1)\big)=\theta_n-\beta_i+\lambda_p+\rho_{i\times p}$ handelt es sich um das wahre Modell, und (b) das Testdesign ist für Items und Itempositionen balanciert. Diese Geringfügigkeit mag ein Grund sein, dass der durch Varianzeinschränkung bedingte Bias in der IRT-Literatur – verglichen mit dem praktisch bedeutenderen Bias aufgrund inadäquater Testdesigns – bisher wenig diskutiert wurde.

Das oben dargestellte Verfahren ist weitgehend analog zu der von Mood (2010) diskutierten Methode der „*y*-standardization". Das Verfahren erlaubt demzufolge Vergleichbarkeit zwischen Modellen, nicht jedoch Vergleichbarkeit zwischen Gruppen. Angenommen, die Itemparameter für Leseaufgaben aus den Large-Scale Assessments zweier hypothetischer Erhebungen der Jahrgänge 2010 und 2012 sollen verglichen werden. Angenommen ferner, dass beidemale bezüglich der Itemposition balancierte Testdesigns genutzt wurden, wobei jedoch Itempositionseffekte im Jahr 2012 eine andere Effektstärke hatten als 2010. Für beide Erhebungen wäre demzufolge ein unterschiedlicher Bias in den Itemparametern des Raschmodells zu erwarten und eine Vergleichbarkeit nicht gegeben. Insofern 2010 und 2012 keine weiteren (unbekannten) Kontexteffekte aufgetreten sind, könnte das oben beschriebene Verfahren der Modellierung eines Rasch- und eines Modells mit Positionseffekten für beide Jahrgänge durchgeführt werden. Für beide Erhebungen würde nun ein unterschiedlicher Quotient aus der Itemvarianz im Rasch- und der Itemvarianz im Modell mit Positionseffekten gefunden werden (im oberen Beispiel war dieses Verhältnis $\frac{1.487}{1.5}\approx 0.9913$. Die Itemparameter des Raschmodells aus den Erhebungen von 2010 und 2012 könnten dann mit dem jeweils erhaltenen Quotienten auf eine neue gemeinsame Metrik transformiert werden, um Vergleichbarkeit herzustellen.

4.2 Identifikation des Bias aufgrund der Varianzeinschränkung

In den vorangegangenen Kapiteln 4 und 4.1 wurden bislang noch keine Aussagen darüber getroffen, wie sich in einem konkreten empirischen Fall erkennen lässt, ob in dem spezifizierten Messmodell bedeutsame Varianzkomponenten nicht mit modelliert wurden. Anders gefragt: Wie lassen sich die praktischen Konsequenzen abschätzen, wenn IRT-Modelle wie das Raschmodell nahezu immer in irgendeiner Weise fehlspezifiziert sind? Woran erkennt man, ob die Fehlspezifizierung wirklich zu einem relevanten Bias führt oder praktisch nicht bedeutsam ist? Wenn balancierte Testdesigns notwendige, aber nicht hinreichende Bedingung für unverfälschte Parameter sind: Wie lässt sich abschätzen, ob die Balancierung in der gewünschten Weise Unverfälschtheit der Modellparameter gewährleistet hat?

Die oben stehenden Fragen zielen weitgehend auf die Frage nach der Passung des Raschmodells auf die empirischen Daten ab. Anders als etwa in linearen Regressionsmodellen oder Strukturgleichungsmodellen gibt es für Modelle der IRT-Familie kein „absolutes“ Kriterium als Maßgabe für die Modellgüte (so wie das R^2 in Regressionsmodellen oder der CFI/TLI bzw. RMSEA in Strukturgleichungsmodellen). Als Gütekriterien werden in IRT-Modellen zumeist die Log-Likelihood (Stoel, Galindo Garre, Dolan & van den Wittenboer, 2006) und daraus abgeleitete Gütemaße wie die Deviance, das *Akaike Information Criterion* (AIC; Akaike, 1974), das *Bayesian Information Criterion* (BIC; Schwarz, 1978), das *Adjusted Bayesian Information Criterion* (ABIC; vgl. Lin & Dayton, 1997) oder das *deviance information criterion* (DIC; Spiegelhalter, Best, Carlin & van der Linde, 2002) genutzt. Hierbei handelt es sich jedoch nur um relative, keine absoluten Gütemaße (Nakagawa & Schielzeth, 2013). AIC, BIC etc. sind stichprobenabhängige Maße und erlauben folglich nur eine Aussage darüber, wie gut ein bestimmtes Modell verglichen mit einem anderen, dazu genesteten Modell *dieselben* Daten beschreibt (Wilson & De Boeck, 2004). Als absolute Werte sind AIC, BIC etc. wertlos, da sie sowohl von der Item- als auch von der Personenstichprobe abhängen (Molenberghs & Verbeke, 2004; Nakagawa & Schielzeth, 2013). In der Praxis wird demzufolge dem präferierten Modell (z. B. Raschmodell, siehe Gleichung 1, Seite 13) ein konkurrierendes, dazu genestetes aber weniger restriktives Modell (z. B. Gleichung 4, Seite 28) gegenüber gestellt und die Passung beider Modelle an die empirischen Daten verglichen. Unternimmt man etwa für die Daten der Teilstichprobe des IQB-Ländervergleichs 2011 in der Primarstufe einen solchen Likelihood-Differenzentest (*likelihood-ratio test*, LRT; Fischer, 1974; McDonald, 2000; Stoel et al., 2006; Verbeke & Molenberghs, 2000), so beschreibt ein Modell, das zusätzlich zu Item- und Personeneffekten Positionseffekte parametrisiert, die Daten signifikant besser (siehe Tabelle A3 im Anhang des ersten Einzelbeitrags, Seite 7). Zu einem ähnlichen Befund würde man anhand der in Tabelle 1 (Seite 37) aufgeführten Gütekriterien kommen: Sowohl AIC als auch BIC weisen für das Modell mit Positionseffekten eine bessere Passung an die empirischen Daten auf.

In vielen praktischen Anwendungen ist ein solcher relativer Modellvergleich jedoch nur eingeschränkt interpretierbar oder nützlich, was im Folgenden näher erläutert werden soll. Genauso wenig wie ein signifikanter Haupteffekt in einer Regression etwas über die praktische Bedeutsamkeit dieses Effekts aussagt (hierzu werden üblicherweise Effektstärkemaße angegeben), sagt diese signifikant bessere Passung des Modells mit Positionseffekten etwas darüber aus, ob dadurch problematische Konsequenzen bezüglich der Validität von Item- oder Personenparametern zu erwarten sind (Sinharay & Haberman, 2014). In Abschnitt 4.1 wurde

demonstriert, dass die praktischen Konsequenzen in diesem konkreten Fall nahezu unbedeutend sind, obwohl der Modellvergleich eine signifikant schlechtere Passung für das Raschmodell ausweist: Der Grund war, dass die hier modellierten Kontexteffekte nur eine sehr geringe Varianz haben. Und selbst falls Kontexteffekte in substanzieller Effektstärker auftreten, müssen daraus keine praktischen Konsequenzen entstehen: Angenommen, ein Kontexteffekt bewirkt in einem fehlspezifizierten Modell einen Bias in den Personen-, nicht jedoch in den Itemparametern. In einer Pilotierungsstudie, wo man vordergründig an der Evaluation von Itemparametern interessiert ist, wäre ein solcher Befund unkritisch; in einer Studie, wo es um die Auswertung von Personenverteilungen geht, wäre ein solcher Befund kritisch. Es müssen also immer die Modellparameter definiert werden, die für die Interpretation der Ergebnisse zentral sind, und es muss anschließend geprüft werden, ob durch gegebenenfalls nicht spezifizierte Kontexteffekte ein Bias *in diesen Modellparametern* auftritt. Die Frage der Validität des Messmodells (*measurement validity*) stellt sich also stets im Zusammenhang mit einem konkreten Anwendungsfall.

Zugleich wäre zur Evaluation der praktischen Bedeutsamkeit des Bias so etwas wie ein „Effektstärkemaß“ vonnöten, was schwierig ist, da es nicht die Bedeutsamkeit eines Kontexteffekts alleine ist, die Aussagen darüber erlaubt, inwieweit die Validität von Modellparametern eingeschränkt ist, sondern zusätzlich die Kombination mehrerer potentieller Kontexteffekte, ihrer Interaktion untereinander und die Interaktion mit Eigenschaften des gewählten Testdesigns. Die Interaktion zweier möglicher Kontexteffekte wird später im zweiten Einzelbeitrag behandelt werden.

Swaminathan et al. (2007) sowie Sinharay und Haberman (2014) schlagen eine andere Möglichkeit vor, um die beiden oben skizzierten Probleme zu lösen. Sie leiten aus bayesianischen Verfahren (Fox, 2010; Fox & Glas, 2001; Kruschke, 2011; Patz & Junker, 1999) eine Methode ab, die Modellpassung bzw. die praktische Bedeutsamkeit einer Nicht-Passung indirekt zu schätzen. Es wird dabei nur das interessierende Modell (in dem hier aufgeführten Beispiel also das Raschmodell) betrachtet. Wie gut dieses parametrisierte Modell die empirischen Daten beschreibt, ist unbekannt. Es kann jedoch geprüft werden, wie Daten aussehen würden, die eine perfekte Passung an genau dieses parametrisierte Modell aufweisen. Hierzu werden beruhend auf den geschätzten Modellparametern künstliche Daten durch Simulation erzeugt. Diese Daten entsprechen hinsichtlich ihrer Item- und Personenstichprobengröße sowie dem Testdesign exakt den empirischen Daten, mit dem einen Unterschied, dass die simulierten Daten eine ideale Passung auf das Raschmodell aufweisen. In der Regel wird dabei nicht einer, sondern es werden viele künstliche Datensätze erzeugt. Die Passung wird bestimmt, indem

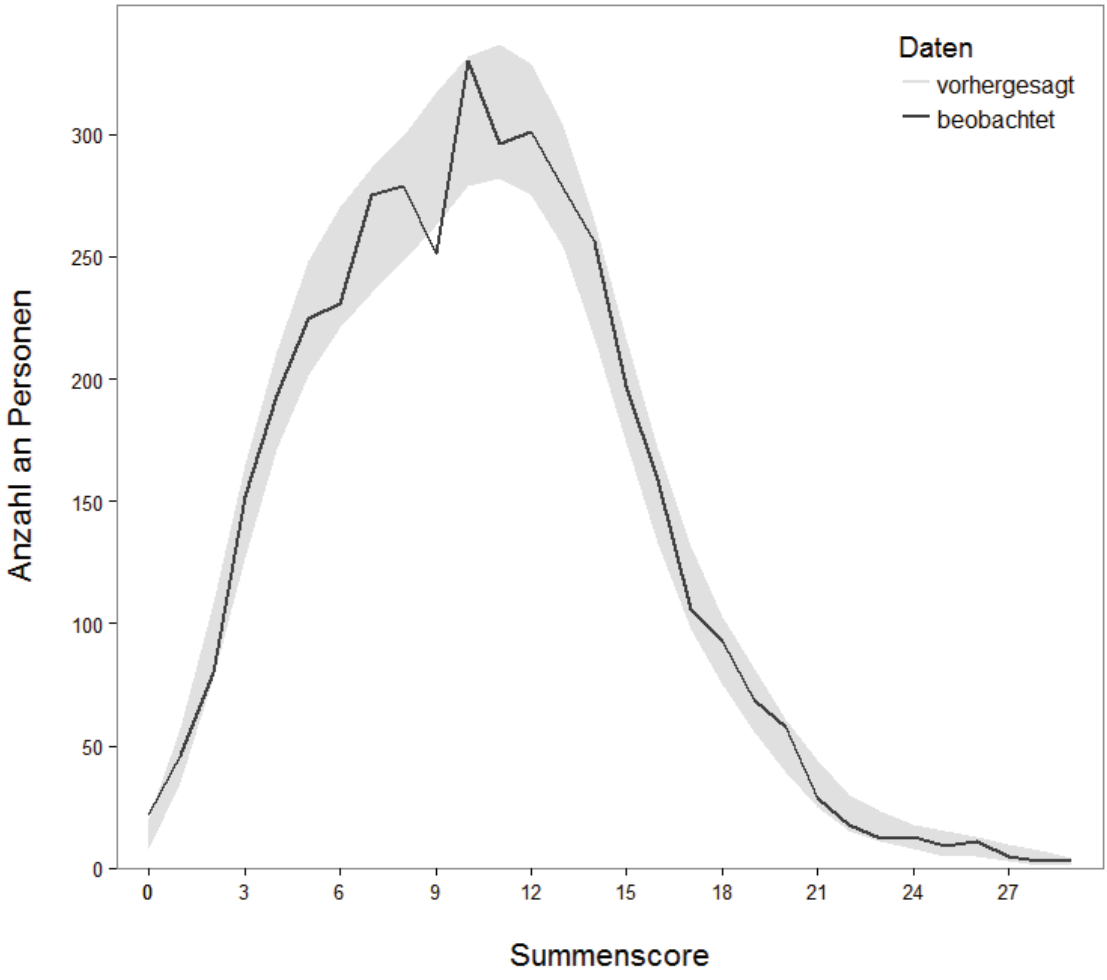
die Unterschiedlichkeit der simulierten Datensätze untereinander ins Verhältnis zu der Unterschiedlichkeit des empirischen Datensatzes von den simulierten Datensätzen gesetzt wird. Das Verfahren beruht auf dem aus bayesianischer Statistik bekannten *Posterior Predictive Model Check* (PPMC; Gelman, Meng & Stern, 1996; Rubin, 1984) der von Sinharay (2005) sowie Sinharay und Johnson (2003) auf IRT-Modelle übertragen wurde. Mit Hilfe der künstlichen Daten wird die Modellpassung indirekt über ein sogenanntes Diskrepanzmaß $D(y^{\text{rep},\,k})$ zwischen den empirisch beobachteten und den künstlichen Daten bestimmt. Streng genommen hat man es hierbei nicht mit einem *Modelltest* oder *Modellvergleich* zu tun, sondern eher mit einem „Datenvergleich", wobei die empirischen Daten mit wiederholten Ziehungen aus der Menge der durch das Modell implizierten Daten verglichen werden. Man vergleicht also nicht Parameter oder Prüfgrößen verschiedener IRT-Modelle, sondern deskriptive Statistiken oder weitere, aus den (empirischen bzw. modellimplizierten) Daten abgeleitete Verteilungen, beispielsweise relative Lösungshäufigkeiten (*p*-Werte) für Items oder Randverteilungen für Personen. Der große Vorteil dabei ist, dass je nach Schwerpunktsetzung entschieden werden kann, ob die praktische Bedeutsamkeit eines gegebenenfalls auftretenden Bias für Items oder Personen oder beides untersucht werden soll.

Sinharay (2006) sowie Sinharay und Haberman (2014) schlagen vor, die Ergebnisse solcher Modellevaluationen grafisch zu veranschaulichen. Ist man an gegebenenfalls auftretenden praktischen Konsequenzen auf der Personenseite interessiert, so lässt sich die empirische Verteilung der Summenwerte je Person mit der durch das Modell implizierten Verteilung vergleichen. Für letztere kann, durch Simulation mehrerer Datensätze aus den Parametern desselben Modells, ein Vertrauensintervall bestimmt werden. Bewegt sich die empirische Verteilung der Summenwerte innerhalb der Grenzen dieses Vertrauensintervalls, so ist ein gegebenenfalls auftretender Bias für die Personenparameter praktisch nicht bedeutsam. Der Nachteil dieses Verfahrens ist, dass im Falle bedeutsamer praktischer Abweichungen der empirischen von der modellimplizierten Verteilung erst einmal nicht klar ist, was die Ursache dieser Abweichungen ist.

Dieses Verfahren soll abermals für empirische Daten des IQB-Ländervergleichs 2011 in der Primarstufe im Fach Deutsch für den Kompetenzbereich Lesen angewendet werden. Die Datengrundlage ist dabei identisch zu dem in Abschnitt 4.1 dargestellten Beispiel. Verwendet wird wiederum die Teilstichprobe von 4000 Schülerinnen und Schülern. Für die Modellierung wurde das einfache Raschmodell aus Gleichung 1 (Seite 13) zugrunde gelegt, dabei wurden sowohl für die Verteilung der Items als auch der Personen nicht-informative Prior-Annahmen definiert, damit die Posteriorverteilung weitestgehend durch die Likelihoodverteilung be-

stimmt ist. Die Analysen wurden in JAGS (Plummer, 2013, 2014) mit Hilfe der Software R (R Core Team, 2014) realisiert. Insgesamt wurden 5000 Iterationen verwendet, wobei die letzten 1000 Iterationen (also 4001 bis 5000) für das PPMC-Verfahren gewählt wurden. Die Ergebnisse sind in Abbildung 1 veranschaulicht.

Abbildung 1: Posterior Predictive Model Check (PPMC) für das Raschmodell und Daten des Hauptdesigns aus dem IQB-Ländervergleich 2011 in der Primarstufe im Fach Deutsch für den Kompetenzbereich Lesen



Der grau schraffierte Bereich kennzeichnet dabei das 95%-Vertrauensintervall, in dem sich die empirische Verteilung der Rohpunktwerte bewegen sollte, sofern es sich bei dem zugrunde gelegten Raschmodell um das „wahre“ Modell handeln würde. Die durchgezogene schwarze Linie bezeichnet die tatsächliche Verteilung der Rohpunktwerte. Eine sichtbare Abweichung der Rohpunktwerte von ihrer erwarteten Verteilung findet sich für den Summenscore von 9 Punkten.

Die hier auftretenden Abweichungen fallen deutlich geringer aus, als sie etwa Sinharay und Haberman (2014) für ein fehlspezifiziertes IRT-Modell in einem Leistungstests (*state test*) finden. Sinharay und Haberman (2014) verwendeten Daten, die auf einem 3PL-Modell beruhten und analysierten sie mithilfe des Raschmodells. Die empirischen Rohpunktwerte waren dabei stärker linkssteil verteilt, als es das Modell implizierte; die Personen erreichten im Mittel weniger Punkte im Test, als bei Gültigkeit des Raschmodell zu erwarten gewesen wäre. Die Ursache der in Abbildung 1 (Seite 42) dargestellten Abweichungen ist jedoch erst einmal nicht eindeutig festzustellen. Um zu prüfen, ob diese hier vorgefundenen Abweichungen darauf zurückgeführt werden können, dass die in den empirischen Daten auftretenden Positionseffekte durch das in der Analyse verwendete Raschmodell nicht berücksichtigt wurden, wurde in einem zweiten Schritt das Modell mit Positionseffekten (Gleichung 4) an denselben Daten wiederum in JAGS mit nicht-informativen Prior-Annahmen geschätzt, und für dieses komplexere Modell ebenfalls ein PPMC durchgeführt. Dabei zeigte sich, dass die Konfidenzbänder aus dem Raschmodell und dem Modell mit Positionseffekten nahezu deckungsgleich waren. Dies kann einerseits als Vorteil aufgefasst werden: Das Testdesign war bezüglich Items und Positionen balanciert; es sollte also bei der Analyse durch das Raschmodell kein Bias aufgrund der Nichtberücksichtigung von Positionseffekten auftreten. Der durch Positionseffekte vermittelte Bias kann also für die Daten des IQB-Ländervergleichs 2011 im Kompetenzbereich Lesen als praktisch nicht bedeutsam eingeschätzt werden. Andererseits legt Abbildung 1 nahe, dass eine Fehlspezifizierung aufgrund eines im Design nicht adäquat berücksichtigten Kontexteffekts vorliegen könnte.

4.3 LRT oder PPMC?

Stellt man diese beiden Methoden – LRT und Verfahren, die auf dem PPMC beruhen – gegenüber, so lässt sich festhalten:

LRTs, die die relative Modellgüte evaluieren, geben keine Auskunft darüber, wie stark die tatsächliche Anpassung des Modells an die empirischen Daten von einer theoretisch „idealen“ Anpassung abweicht (Dziak, Coffman, Lanza & Li, 2012; Molenberghs & Verbeke, 2004). Sie geben keine Auskunft über die praktische Bedeutsamkeit einer Fehlspezifizierung des Modells.

PPMC-Verfahren geben keine Auskunft über die Ursache einer möglichen Nicht-Anpassung. Sie evaluieren im strengen Sinne auch nicht ein Modell oder die Passung eines Modells, sondern bieten lediglich eine Möglichkeit, aus den durch das Modell implizierten Daten abgeleitete Verteilungen zu vergleichen (Sinharay & Haberman, 2014).

Das Verfahren der Wahl könnte daher sein, beide Methoden zu kombinieren: Der PPMC kann verwendet werden, um zu prüfen, ob das zugrunde gelegte Modell bezüglich der gewünschten Inferenz in praktisch bedeutsamer Weise fehlspezifiziert ist. Falls das der Fall ist, könnte die Ursache dafür näher identifiziert werden, indem konkrete Hypothesen formuliert und verschiedene konkurrierende Modelle geschätzt und deren Anpassung über Log-Likelihood-Tests gegen das ursprüngliche Modell verglichen werden. Das erste Verfahren wäre eher explorativ, das zweite – im inferenzstatistischen Sinne – konfirmatorisch. Zudem könnte das Verfahren fortgesetzt werden, indem das durch den Log-Likelihood-Test präferierte Modell nun seinerseits in einem PPMC auf praktisch bedeutsame Fehlspezifizierungen geprüft wird.

Dabei ist jedoch zu berücksichtigen, dass das Testdesign für sämtliche Modelle, die in einem solchen schrittweisen Verfahren geprüft werden, adäquat ist. Ein zentraler Befund aus dem ersten Einzelbeitrag „Modeling Item Position Effects Using Generalized Linear Mixed Models“ war, dass die Prüfstatistik des Log-Likelihood-Tests nur dann unverzerrt ist, wenn das Testdesign bezüglich der durch das Modell parametrisierten Faktoren balanciert ist.

4.4 Zusammenfassende Einschätzung der praktischen Bedeutsamkeit des Bias

In dem ersten Einzelbeitrag der Dissertation und dem Kapitel 4 der Rahmung wurde deutlich, dass Kontexteffekte zu verzerrten Schätzungen von Itemparametern führen können. Zwei mögliche Ursachen, bedingt durch die Kontexteffekte, kommen dabei als Ursache des Bias infrage:

1. Das Testdesign ist nicht ausbalanciert.
2. Unterkomplexe IRT-Modelle führen zu einer Varianzeinschränkung und damit zu einer Veränderung der Metrik der Skala.

Die erste Ursache kann durch ein entsprechend angepasstes Testdesign eliminiert werden, die zweite jedoch nicht. Was würde ein solcher Bias, übertragen auf die Metrik der Bildungsstandards bedeuten? Betrachtet man das für den Ländervergleich 2011 gewählte Design (vollständig balanciert) und die empirisch ermittelte Stärke der Positionseffekte (mittelstarke, nicht-lineare Effekte; siehe Tabelle A2 im Anhang des ersten Einzelbeitrags, Seite 6), so wäre hierfür ein mittlerer Bias von 0.077 zu erwarten (siehe Tabelle 3 des ersten Einzelbeitrags, Seite 544, Zeile 8). Um diesen Wert auf die Metrik der Bildungsstandards zu übertragen, setzt man ihn ins Verhältnis der Personenvarianz der Schülerinnen und Schüler im Ländervergleich 2011 der vierten Jahrgangsstufe in Deutschland von 1.035 (Weirich & Pant, 2014) und erhält

einen Wert von $100 \cdot \frac{0.077}{\sqrt{1.035}} \approx 7.19$ Punkten. Diese Abweichung ist im Betrag doppelt so groß wie der mittlere Standardfehler für Itemparameter des Kompetenzbereichs Lesen im Ländervergleich 2011. Sie kann nach oben oder nach unten auftreten und möglicherweise bedeuten, dass Items einer falschen Kompetenzstufe zugewiesen werden würden. Berücksichtigt man, dass Items auf Kompetenzstufe 1 nur durch eine Abweichung nach oben, Items auf Kompetenzstufe 5 nur durch eine Abweichung nach unten, Items auf den Kompetenzstufen 2, 3 und 4 durch Abweichungen sowohl nach oben oder nach unten einer falschen Kompetenzstufe zugewiesen werden können, so würde ein mittlerer Bias von 7.19 Punkten bedeuten, dass aufgrund von Positionseffekten durchschnittlich $0.2 \cdot \left[3 \cdot \left(1 - \frac{75 - 7.19}{75} \right) + 2 \cdot \left(1 - \frac{75 - 0.5 \cdot 7.19}{75} \right) \right] \approx 7.67$ Prozent der Items einer falschen Kompetenzstufe zugewiesen werden würden.

Da die Parameter auf der Personenebene stets nur in Relation zu den Itemparametern bestimmt werden, könnte der oben beschriebene mittlere Bias von 7.19 Punkten in ähnlicher Weise die Schätzung von Personenverteilungen beeinflussen. Um die Leistung zweier verschiedener Personenpopulationen anhand einer Teilstichprobe gemeinsamer Items zu vergleichen, werden üblicherweise in einem *common-item nonequivalent groups equating design* (Kolen & Brennan, 2004; von Davier, A. et al., 2008) die Parameter der in beiden Gruppen gemeinsam auftretenden Items auf gemeinsame Werte oder feste Referenzwerte fixiert. Sofern nun beide Gruppen in unterschiedlicher Weise von Positionseffekten betroffen sind, wäre in beiden Gruppen ein unterschiedlicher mittlerer Bias in den Itemparametern zu erwarten, und die Differenz dieses Bias würde als Bias der geschätzten Differenz beider Personenpopulationen wirken. Die Auswirkungen von Kontexteffekten im Allgemeinen und Positionseffekten im Besonderen *auf die Personenebene* sind also stets nur in Abhängigkeit zu den entstehenden Auswirkungen auf Itemebene zu beurteilen. Dies hat auch damit zu tun, dass in den Large-Scale Assessments zugrunde liegenden Messmodellen die Personenparameter in einem zweiten Schritt bedingt auf fixierte Itemparameter geschätzt werden (Adams & Wu, 2007; von Davier, M. et al., 2007, siehe auch Kapitel 7.1)

Für das hier aufgeführte Beispiel wären etwa keine *über den Bias auf Itemebene hinausgehenden* Auswirkungen für die Personenebene zu befürchten: Aus Tabelle 1 (Seite 37) geht hervor, dass die Varianz der Personen kaum durch die Nichtberücksichtigung von Positionseffekten beeinflusst wird (Varianz von 1.019, bzw. 1.020). Auch das Intercept, das hier als die Differenz zwischen mittlerer Itemschwierigkeit und mittlerer Personenfähigkeit interpretiert

werden kann, bleibt für das Modell mit und das Modell ohne Parametrisierung von Positionseffekten nahezu unverändert (0.562 bzw. 0.560).

Die hier getroffenen Einschätzungen gelten jedoch nur für aggregierte Statistiken (also Mittelwert und Varianz der Gesamtheit aller Personen) und auch nur im Rahmen von Large-Scale Assessments. Anders sähe es etwa aus, wenn nach dem Bias für individuelle Personenfähigkeitswerte gefragt werden würde, oder wenn die hier auftretenden Phänomene im Rahmen von computerisierten adaptiven Tests (CAT) (Kingston & Dorans, 1984; Wainer & Kiely, 1987) beurteilt werden würden. Das Prinzip der Ausbalancierung (siehe Kapitel 2.7) eliminiert nicht die Wirkung von Kontexteffekten, sondern gewährleistet lediglich, dass die interessierenden Parameter der Messung (d. h., Items und Personen) *im Mittel* in gleicher Weise davon beeinflusst sind. Diese Methoden können daher nicht in derselben Weise genutzt werden, wenn das Ziel der Messung darin besteht, individuelle Kennwerte (z. B. für die Fähigkeit der Testpersonen) zu gewinnen.

5. Kontexteffekte auf der Personenseite

In den bisherigen Kapiteln wurden Kontexteffekte auf der Seite der Items oder des Tests behandelt, meist am Beispiel von Positionseffekten: Die Position eines Items beeinflusst die Schwierigkeit – oder den Effekt – des Items. Obwohl diese Kontexteffekte auf der Itemseite auftreten, können auch Parameter auf der Personenseite dadurch verfälscht werden. Item- und Personenparameter können stets nur in Relation zueinander geschätzt werden, wobei entweder der Mittelwert der Item- oder der Mittelwert der Personenpopulation auf Null gesetzt wird (Embretson & Reise, 2000). In praktischen Anwendungen und Large-Scale Assessments ist meist Letzteres der Fall (Allen et al., 2001; Frey, Carstensen, Walter, Rönnebeck & Gomolka, 2008; Weirich, Haag & Roppelt, 2012; vgl. auch Kapitel 7.1). Für Positionseffekte bedeutet dies, dass es letztlich immer eine Frage der Interpretation ist, ob durch Positionseffekte die Items schwerer *oder* die Personen weniger leistungsfähig werden (vgl. Hartig & Buchholz, 2012). Dies gilt auch für andere Arten von Kontexteffekten auf Item- oder Testebene, etwa Testheft- oder Bookleteffekte (Hecht, Weirich, Siegle & Frey, 2014). Kontexteffekte können jedoch auch auf der Personenseite auftreten.

Kontexteffekte auf der Personenseite dürfen dabei nicht mit Effekten der Personen selbst verwechselt werden. Wenn Mädchen in einem Test zur Lesefähigkeit besser abschneiden als Jungen, ist das kein Kontexteffekt des Geschlechts g; es bedeutet lediglich, dass es sich bei der Verteilung von θ nicht um eine homogene Normalverteilung, sondern um eine auf g be-

dingte Mischverteilung handelt. Kontexteffekte auf der Personenseite kommen erst dann zustande, wenn durch Eigenschaften des Tests bestimmte Eigenschaften auf der Personenseite differentiell beeinflusst werden, und wenn diese Personeneigenschaften in Zusammenhang mit dem zu messenden Konstrukt stehen (Brennan, 1992; Marsh, 1984; Messick, 1984; vgl. auch Kapitel 2.3). Ein Beispiel dafür ist die Testteilnahmemotivation (Baumert & Demmrich, 2001; Eklöf, 2010; Penk, Pöhlmann & Roppelt, 2014; Wise & DeMars, 2005; Wise & Kong, 2005; Wise & Smith, 2011). Wenn die Motivation der Testteilnehmer einen signifikanten Einfluss auf die Lösungswahrscheinlichkeit $P(X_{nip} = 1)$ hat und die Motivation der Testteilnehmer aufgrund verschiedener Testadministrationsbedingungen variiert (z. B. wenn die Testergebnisse in einigen Klassen benotet werden, in anderen nicht), kann das unerwünschte Kontexteffekte zur Folge haben. Analog zu der Position eines Items würde die Motivation der Testteilnehmer nicht nur nicht konstant gehalten, sondern systematisch zwischen Testteilnehmern variieren.

Möglich (und weit schwieriger zu kontrollieren) sind Effekte, bei denen die laufende Testbearbeitung ihrerseits die Motivation beeinflusst. Leistungsschwache Schülerinnen und Schüler könnten im Verlauf des Tests aufgrund der subjektiv wahrgenommenen Überforderung stärker in ihrer Motivation nachlassen als leistungsstarke Schülerinnen und Schüler. Der Testwert würde die Leistung von leistungsschwachen Schülerinnen und Schülern folglich unterschätzen. Aus diesem Grunde ist man bestrebt, die Schwierigkeit der Testaufgaben möglichst gut an die vermutete Fähigkeit der Testteilnehmer anzugleichen (Asseburg & Frey, 2013; Van der Linden et al., 2004), obschon ein solches Vorgehen problematisch sein kann, wenn Testhefteffekte auftreten, und die Testhefte nicht zufällig unter den Testteilnehmern verteilt sind (Hecht et al., 2014).

Wenn in einem Test mit Kontexteffekten auf der Personenseite gerechnet werden muss, so gilt für das Testdesign dasselbe, was auch schon für Kontexteffekte auf der Itemseite eingewendet wurde: Um eine unverzerrte Schätzung von Personenparametern zu gewährleisten, müssen die Testbedingungen bezüglich der Personenvariablen entweder standardisiert oder das Testdesign bezüglich dieser Personenvariablen ausbalanciert sein. Standardisierte Testbedingungen, wie sie etwa in Large-Scale Assessments realisiert werden, bewirken jedoch nicht immer auch, dass alle möglichen Kontextvariablen auf der Personenseite ihrerseits standardisiert sind. Am Beispiel der Testteilnahmemotivation wird das deutlich:

Large-Scale Assessments haben häufig für die Testteilnehmer keine praktischen Konsequenzen; die Testergebnisse werden nicht benotet, und es findet keine von der geleisteten Anstrengung abhängige Aufwandsentschädigung statt. Üblicherweise werden die Testteilnehmer

nur appellativ oder durch kleine Gegenleistungen (*incentives*) ermutigt, sich in dem Test anzustrengen. Solche als *low stakes* bezeichnete Testbedingungen sollten für alle Testteilnehmer in derselben Weise gelten. Da jedoch die Testergebnisse auf bildungsadministrativer Ebene einen großen Stellenwert genießen, kann nicht ausgeschlossen werden, dass die unternommenen Anstrengungen zur Gewährleistung einer möglichst hohen Testteilnahmemotivation systematisch *zwischen den Gruppen* variieren, die beispielsweise in einem Test Gegenstand des Vergleichs sind (z. B. Personen verschiedener Bundesländer im Ländervergleich). Hier könnte also eine im Mittel zwischen Gruppen variierende Testteilnahmemotivation als Kontexteffekt zu verzerrten Ergebnissen führen.

Selbst wenn die Testbedingungen zwischen den zu vergleichenden Gruppen exakt identisch sind, ist damit zu rechnen – analog zu dem in Abschnitt 2.7 genannten Beispiel zu Positionseffekten –, dass die Varianz der Testteilnahmemotivation nicht Null ist. In dem zweiten Einzelbeitrag wird daher der Frage nachgegangen, ob eine potentielle wechselseitige Abhängigkeit zweier Kontexteffekte (Positionseffekte auf der Itemseite, Motivationseffekte auf der Personenseite) besteht, und wie sich diese Abhängigkeit auf die gezeigte Testleistung auswirkt.

6. Zusammenfassung des zweiten Einzelbeitrags: Item Position Effects are Moderated by Changes in Test-Taking Effort

Der zweite Beitrag wurde am 25.11.2014 bei der Zeitschrift *Journal of Educational Measurement* eingereicht. Die vorläufige Referenz lautet:
Weirich, S., Penk, C., Hecht, M., Roppelt, A., & Böhme, K. (under review). Item position effects are moderated by changes in test-taking effort. *Journal of Educational Measurement.*

Der erste Beitrag behandelte die Frage, wie (unidimensionale) Itempositionseffekte adäquat modelliert werden können. Daran anschließend, wurde in dem zweiten Einzelbeitrag der Ursache von Itempositionseffekten nachgegangen. Es sollten Faktoren gesucht werden, die die Stärke der Positionseffekte moderieren. In Anlehnung an Debeer und Janssen (2013) wurde die Anstrengungsbereitschaft (*test-taking effort*) als ein solcher Faktor vermutet. Zur Modellierung wurden Daten des IQB-Ländervergleichs 2012 in den Naturwissenschaften Physik, Chemie und Biologie für die Jahrgangsstufe 9 (Pant et al., 2013) verwendet. Verschiedene Skalen zur Testteilnahmemotivation (Anstrengungsbereitschaft, Motivation, Testängstlichkeit etc.) wurden im Ländervergleich zu drei Messzeitpunkten (vor dem Test, nach der Hälfte der Testbearbeitung, nach dem Test) erhoben. Dies ermöglichte es, die Veränderung der Testteilnahmemotivation in einem „Mikro-Längsschnitt“ (*micro longitudinal design*) zu modellieren.

Für den zweiten Einzelbeitrag wurde dabei lediglich eine Skala dieses mehrdimensionalen Konstrukts – die Anstrengungsbereitschaft – betrachtet.

Im ersten Schritt wurde die Veränderung der Anstrengungsbereitschaft über den Testverlauf in einem *latent growth curve model* spezifiziert. Konkret wurde ein *curves-of-factor model* (Duncan, Duncan & Strycker, 2006) in Mplus, Version 7.11 (Muthén & Muthén, 1998-2012) modelliert. Faktorladungen und Intercepts wurden als unveränderlich über die Messzeitpunkte angenommen. Es zeigte sich mit einem standardisierten Regressionskoeffizienten von –0,76 ein beträchtlicher Abfall der Anstrengungsbereitschaft über den Testverlauf.

In einem zweiten Schritt wurden nichtlineare Itempositionseffekte mehrdimensional in einem allgemeinen linearen gemischten Modell parametrisiert. Erwartungsgemäß zeigte sich ein negativer Effekt (die Lösungshäufigkeit für ein Item nimmt mit steigender Position ab), wobei die Effektstärke für Position 4 geringer ausfällt als für Position 3. Die vermutete Ursache dafür ist die zehnminütige Unterbrechung des Tests nach 60 Minuten Bearbeitungszeit, die den Schülerinnen und Schülern Gelegenheit zur Erholung bietet, wodurch der Positionseffekt unmittelbar nach der Pause geringer ausfällt. Die Positionseffekte waren heterogen über die Personen; das bedeutet, Positionseffekte wirken nicht für alle Personen in derselben Weise. Dieser Befund macht es plausibel, nach weiteren Faktoren auf der Personenseite zu suchen, die diese differentiellen Positionseffekte moderieren könnten.

Im dritten Schritt zeigte sich schließlich, dass Positionseffekte stärker ausgeprägt sind, wenn die Personen zu Beginn eine geringere Anstrengungsbereitschaft angaben. Darüber hinaus waren Positionseffekte für solche Personen stärker ausgeprägt, die über den Testverlauf einen stärkeren Abfall in ihrer Anstrengungsbereitschaft zeigten. Die Befunde zeigen, dass Kontexteffekte nicht homogen über alle Personen und isoliert von anderen Effekten betrachtet werden können; sie interagieren vielmehr mit Variablen auf der Personenebene und sogar mit anderen Kontexteffekten. Diese Tatsache macht es sehr schwer, für einen bestimmten Einzelfall die konkreten Konsequenzen (z. B. bezüglich eines zu erwartenden Bias in den Parametern) abzuschätzen. Dies gilt auch für die inhaltliche Interpretation der Ergebnisse: Der substanzielle Effekt der Anstrengungsbereitschaft auf die Wahrscheinlichkeit einer korrekten Itemantwort (siehe Tabelle B3 des dritten Einzelbeitags, Seite 84) kann so verstanden werden, dass Large-Scale Assessments nicht (ausschließlich) die Fähigkeiten messen, die Schülerinnen und Schüler beherrschen, sondern die Fähigkeiten, die sie zu demonstrieren gewillt sind.

7. Missing Data als Kontexteffekt?

Der dritte und letzte Einzelbeitrag befasst sich mit dem Problem fehlender Werte (*missing data*) auf Hintergrundvariablen. Üblicherweise werden fehlende Werte auf Hintergrundvariablen nicht als Kontexteffekte verstanden, obschon sie – in der in Abschnitt 2.3 definierten Weise – durch Eigenschaften des Tests zwar nicht zustande kommen, aber beeinflusst werden können. Der Kontext hierbei wären (möglicherweise unbeobachtete) Variablen, die die Wahrscheinlichkeit fehlender Werte vorhersagen. Das soll im Folgenden näher ausgeführt werden.

Grundsätzlich können fehlende Werte in Large-Scale Assessments in dreierlei Weise auftreten.

1. Das zu messende Konstrukt stellt – allein dadurch, dass es als latent, also nicht direkt beobachtbar definiert wird – eine Variable dar, auf der jede Person einen fehlenden Wert hat.
2. Da in Large-Scale Assessments nicht alle Items von allen Personen bearbeitet werden können, wird bereits durch das Testdesign ein Muster (*pattern*) fehlender Werte definiert.
3. Wenn Schülerinnen und Schüler intentional Testaufgaben oder Fragebogenitems, die ihnen zur Beantwortung vorgelegt werden, nicht bearbeiten, entstehen fehlende Werte, die nicht beabsichtigt sind.

Der erste Fall fehlender Werte ist eine direkte Konsequenz des Messmodells, und der zweite Fall fehlender Werte eine direkte Konsequenz des gewählten Testdesigns. In dem dritten Einzelbeitrag wird eingangs ausgeführt, dass für beide Fälle geeignete Methoden etabliert sind, die es ermöglichen, unverzerrte Parameterschätzer zu gewinnen.

Der dritte Fall fehlender Werte kann dagegen unter Umständen zu verzerrten Schätzungen auf der Personenebene führen (Rutkowski, 2011). Hierzu soll auf eine Besonderheit der praktischen Anwendung des Raschmodells in großen Schulleistungsstudien eingegangen werden.

7.1 Das Raschmodell in Large-Scale Assessments: zwei Modelle in einer Abhängigkeitsstruktur

In Abschnitt 2.1 wurde gesagt, dass sich das Raschmodell aus den drei Komponenten Modellgleichung, Transformationsfunktion und der Zufallskomponente zusammensetzt. Für die Modellgleichung (Gleichung 1) bedeutet das, dass für jede Person und jedes Item ein Parameter geschätzt werden müsste. Die Tatsache, dass das Modell für größere Stichproben eine größere Anzahl an Parametern benötigt, führt zu inkonsistenten Parameter-Schätzungen (Embretson &

Reise, 2000; Tuerlinckx et al., 2004; Wilson & De Boeck, 2004). Darüber hinaus ist man in Large-Scale Assessments an *individuellen* Personenparametern meist nicht interessiert, es geht vielmehr um die Schätzung (bedingter) Populationsverteilungen. Aus beiden genannten Gründen ist es sinnvoll, das Raschmodell insofern zu variieren, als dass den drei Komponenten eine weitere Annahme hinzugefügt wird, und zwar über die Verteilung des Personenparameters θ_n. Es wird also nicht mehr ein Parameter je Person, sondern nur ein Lageparameter (Mittelwert μ) und ein Dispersionsparameter (Varianz σ_θ^2) für die gesamte Population bestimmt. Im einfachsten Fall wird dabei eine univariate Normalverteilung angenommen: $\theta_n \sim N\left(\mu, \sigma_\theta^2\right)$. Jede einzelne Person der Stichprobe wird folglich nur als eine Zufallsziehung aus der durch die beiden Parameter μ und σ_θ^2 definierten Population betrachtet. Da die Effekte der Personen als zufällig (*random*) und nur die Effekte der Items als fixiert (*fixed*) betrachtet werden, wird diese Formulierung des Raschmodells auch als random person–fixed items (RPFI) bezeichnet (De Boeck, 2008). In technischer Hinsicht, das heißt, was den der Schätzung zugrundeliegenden Algorithmus angeht, entspricht dieses Modell der *Marginal maximum likelihood* (MML) Methode (Embretson & Reise, 2000; Tuerlinckx et al., 2004). Andere Methoden sind etwa *Joint Maximum likelihood* (JML), was einer Formulierung des Raschmodells als fixed person–fixed item (FPFI) entspricht.

Anders als bei linearen Messmodellen können die Parameter log-linearer latenter Messmodelle nicht analytisch (etwa über die Methode der kleinsten Quadrate) bestimmt werden. Die verwendeten Approximierungen (Gauss-Hermite, Laplace, Markov Chain Monte Carlo) nutzen daher meist zusätzliche Annahmen, die die formale Definition des Raschmodells implizit ergänzen. Modellformulierung und Schätzalgorithmus sind also nicht unabhängig voneinander.

Für die praktische Anwendung in Large-Scale Assessments sind diese zusätzlichen, im Grunde restriktiven Annahmen über die Verteilung der Population jedoch von Vorteil. Ist man beispielsweise daran interessiert, nicht nur Mittelwert und Standardabweichung der gesamten Population für ein bestimmtes univariates Konstrukt (z. B. mathematische Kompetenz) zu schätzen, sondern darüber hinaus den Zusammenhang von mathematischer und sprachlicher Kompetenz, kann die für RPFI getroffene Annahme über die Verteilung von θ_n erweitert werden. Man würde nicht länger eine univariate Annahme über die Verteilung des Merkmals in der Population treffen, sondern eine multivariat normalverteilte (*MVN*) Population unterstellen: $\theta_n \sim MVN\left(0, \Sigma\right)$, mit $\Sigma = \begin{pmatrix} \sigma_{MM}^2 & \\ \sigma_{MS} & \sigma_{SS}^2 \end{pmatrix}$. Hier würde nicht mehr nur die Varianz der Po-

pulation, sondern die Varianz für mathematische (σ^2_{MM}) und sprachliche Kompetenz (σ^2_{SS}) sowie ihre Kovarianz (σ_{MS}) geschätzt. Wenn darüber hinaus angenommen wird, diese Verteilung hänge von weiteren Variablen auf der Personenseite ab (etwa dem Geschlecht *Z*), so kann statt einer multivariaten Normalverteilung eine multivariate Mischverteilung spezifiziert werden: $\theta_n \sim MVN(\boldsymbol{\mu}, \Sigma)$, mit $\Sigma = \begin{pmatrix} \sigma^2_{MM} & \\ \sigma_{MS} & \sigma^2_{SS} \end{pmatrix}$. $\boldsymbol{\mu}$ ist dabei ein Vektor mit zwei Elementen für die beiden Mittelwerte von mathematischer und sprachlicher Kompetenz. Des weiteren kann $\boldsymbol{\mu}$ dabei für beide Kompetenzbereiche linear zerlegt und in einer Regression durch die Geschlechtsvariable *Z* vorhergesagt werden.

Die Möglichkeit, latente univariate oder multivariate Verteilungen in der Population bedingt auf einen Vektor von Kovariaten zu schätzen, macht man sich in latenten Hintergrundmodellen zunutze, die prinzipiell sowohl für ein- als auch für mehrdimensionale 1PL-, 2PL- oder 3PL-Modelle angewendet werden können. In Large-Scale Assessments ist die Bestimmung von Fähigkeitsunterschieden in Teilpopulationen (z. B. die verschiedenen Länder der Bundesrepublik Deutschland, Kinder aus Familien mit einem hohen vs. einem niedrigen sozio-ökonomischen Status) zentral. RPFI-Raschmodelle (bzw. MML-formulierte Raschmodelle) mit latentem Hintergrundmodell sind hierfür geeignet. Für den unidimensionalen Fall wäre dann Gleichung 1 (Seite 13) durch die unten dargestellte Gleichung 7 zu erweitern:

$$\text{logit}\left(P(X_{ni}=1)\right) = \theta_n - \beta_i, \text{ und}$$

$$\theta_n = \mathbf{Y_n}\boldsymbol{\beta} + E_n, \text{ mit } E_n \sim N\left(0, \sigma^2_\theta\right). \qquad (7)$$

Die Definition von θ_n wird erweitert, und θ_n nunmehr linear zerlegt in einen Vektor aus Parametern $\boldsymbol{\beta}$ für den Vektor aus Hintergrundvariablen $\mathbf{Y}$, sowie die normalverteilte Fehlerkomponente E_n.

7.2 Bedeutung latenter Hintergrundmodelle

Die Bedeutung latenter Hintergrundmodelle wurde in der Literatur bereits vielfach diskutiert (Adams & Wu, 2007; Frey et al., 2008; Mislevy, Beaton, Kaplan & Sheehan, 1992; von Davier, M., Gonzalez & Mislevy, 2009; von Davier, M. et al., 2007). Es soll daher an dieser Stelle lediglich auf einige Besonderheiten in der praktischen Umsetzung in Large-Scale Assessments eingegangen werden.

Die in Abschnitt 7.1 beschriebene Erweiterung des Raschmodells um das latente Hintergrundmodell hat praktische Auswirkungen auf die Schätzung der Parameter. Wie in Abschnitt

5 erläutert, können die Schwierigkeit von Items und die Fähigkeit von Personen immer nur in Relation zueinander gemessen werden. Damit das Modell identifiziert ist, wird dabei i. d. R. per Definition der Mittelwert der Personenpopulation auf 0 fixiert. Im Falle der Verwendung eines Hintergrundmodells wird jedoch das Intercept der Modellgleichung im Hintergrundmodell (siehe Gleichung 7) auf 0 fixiert. Das bedeutet allerdings: Wenn sich das Skalenniveau der Variablen im Hintergrundmodell ändert, können sich das Intercept und demzufolge auch die Itemparameter ändern. Gängige Praxis ist es daher, in einem ersten Schritt zunächst die Items ohne Hintergrundmodell zu kalibrieren und in einem zweiten Schritt die Parameter des Hintergrundmodells zu schätzen, wobei die Itemparameter auf die im ersten Schritt geschätzten Werte fixiert werden (Frey et al., 2008; von Davier, M. et al., 2007). Das bedeutet: Um zu Schätzern für die Parameter des Hintergrundmodells zu gelangen, wird nicht ein, sondern es werden *zwei* Modelle geschätzt, die in einer Abhängigkeitsbeziehung zueinander stehen: das zweite Modell wird bedingt auf die Parameter des ersten geschätzt. Diese Praxis wird insbesondere von Vertretern bayesianischer Verfahren kritisiert, da die Unsicherheit bei der Schätzung der Itemparameter im ersten Schritt für die Parameterschätzung im zweiten Schritt nicht berücksichtigt wird (Patz & Junker, 1999).

Diese Abhängigkeitsbeziehung bedeutet zugleich: Jeder Kontexteffekt, der einen Bias für Itemparameter bewirkt, bewirkt damit auch einen Bias für Personenparameter, da diese in Abhängigkeit zu den Itemparametern geschätzt werden. Das Umgekehrte gilt jedoch nicht: das von Rutkowski (2011) behandelte Problem fehlender Werte im Hintergrundmodell auf die Güte der geschätzten Modellparameter betrifft ausschließlich Variablen in Gleichung 7; das bedeutet, ein eventueller Bias kann sich nur auf Personen-, nicht aber auf Itemparameter auswirken.

8. Zusammenfassung des dritten Einzelbeitrags: Nested multiple imputation in large-scale assessment

Dieser Beitrag ist in der Zeitschrift *Large-Scale Assessments in Education* erschienen. Die Referenz lautet:

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T. & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-Scale Assessments in Education, 2*(9), 1-18.

In dem dritten Einzelbeitrag wird ein Bezug zu einem Befund von Rutkowski (2011) hergestellt: Ein substanzieller Anteil fehlender Werte auf Hintergrundvariablen kann zu verzerrten Parameterschätzern auf Personenebene führen, wenn diese fehlenden Werte – wie bisher

in Large-Scale Assessments üblich – lediglich über Dummyvariablen indiziert werden. In dem Beitrag wird daher das Prinzip multipler Imputation (Graham, 2009; Little, 1992; Little & Rubin, 1987; Rubin, 1987; Schafer & Graham, 2002) auf latente Hintergrundmodelle übertragen. Dies bedeutet, dass zu den in Kapitel 7 angesprochenen zwei Modellen (Kalibrierungsmodell und Hintergrundmodell bzw. *conditioning model*) für die Bestimmung der Parameter des Hintergrundmodells nun noch ein weiteres, drittes Modell dazu kommt, das Imputationsmodell für die fehlenden Werten auf den Hintergrundvariablen. Die Gewinnung personenspezifischer Werte (*plausible values*) setzt also zwei Imputationsverfahren in einer Abhängigkeitsstruktur voraus, was in der Literatur als genestete Imputation (Rubin, 2003) oder *two-stage multiple imputation* (Harel, 2007; Harel & Schafer, 2003; Reiter & Drechsler, 2007; Reiter & Raghunathan, 2007) beschrieben wird.

In dem Einzelbeitrag wird im Rahmen einer Simulationsstudie geprüft, ob dieses Verfahren das Problem des von Rutkowski (2011) beschriebenen Bias aufgrund fehlender Werte auf Hintergrundvariablen beheben kann. Dabei wurden zwei verschiedene Methoden untersucht, wobei das erste Mal jeder fehlende Wert auf einer Hintergrundvarablen nur durch eine Imputation (*single + multiple imputation, SMI*) und das zweite Mal jeder fehlende Wert auf einer Hintergrundvariablen durch fünf Imputationen (*multiple + multiple imputation, MMI*) ersetzt wurde. Um die Bedingungen empirischer Assessments möglichst realistisch abzubilden, wurden in der Simulation sowohl eine Abhängigkeit der tatsächlichen Testleistung von den Hintergrundvariablen, als auch eine Abhängigkeit der Testleistung von der Wahrscheinlichkeit fehlender Werte auf den Hintergrundvariablen jeweils in zwei verschiedenen Effektstärken simuliert.

Im Ergebnis zeigten sich bezüglich des Bias keine bedeutsamen Unterschiede zwischen SMI und MMI. Der mittlere Bias war in beiden Fällen nahezu Null. Im Falle des gemeinsamen Auftretens verschiedener kritischer Simulationsbedingungen (hohe Abhängigkeit der Testleistung von den Werten auf Hintergrundvariablen sowie der Wahrscheinlichkeit fehlender Werte auf Hintergrundvariablen *und* hoher Anteil fehlender Werte auf Hintergrundvariablen) stieg der Bias bis auf Werte von 0,145 im Betrag. Um diesen Wert auf die Metrik der Bildungsstandards zu übertragen, setzt man ihn ins Verhältnis der Personenvarianz der Schülerinnen und Schüler im Ländervergleich 2011 der vierten Jahrgangsstufe in Deutschland von 1.035 (Weirich & Pant, 2014) und erhält einen Wert von $100 \cdot \frac{0.145}{\sqrt{1.035}} \approx 14.25$ Punkten. Analog zu dem Beispiel aus Kapitel 4.4 würde ein mittlerer Bias von 14.25 Punkten bedeuten,

dass durchschnittlich $100 \cdot 0.2 \cdot \left[3 \cdot \left(1 - \frac{75 - 14.25}{75} \right) + 2 \cdot \left(1 - \frac{75 - 0.5 \cdot 14.25}{75} \right) \right] \approx 15.2$ Prozent der Personen einer falschen Kompetenzstufe zugewiesen werden würden.

Bezüglich des RMSE und der *recovery proportion* fanden sich leichte Vorteile von MMI gegenüber SMI. Die Unterschiede zwischen SMI und MMI sollten sich theoretisch lediglich auf die Bestimmung der Standardfehler und damit die Breite der Konfidenzintervalle für die Parameterschätzer auswirken. Praktisch war der Anteil derjenigen Simulationsbedingungen, in denen das Konfidenzintervall (etwa der geschätzten Mittelwertsdifferenz) den wahren Wert, also die tatsächliche Mittelwertsdifferenz einschloss, für MMI größer als für SMI.

9. Diskussion und abschließende Bewertung der Ergebnisse

Kontexteffekte stellen im Rahmen von auf der Item-Response-Theorie beruhenden Analysen ein besonderes Problem dar. Der Grund ist, dass sowohl die Effekte selbst, als auch ihre Auswirkungen oft nicht direkt beobachtet und/oder identifiziert werden können. So ist es zwar möglich, über das sogenannte „Infit"-Kriterium (Adams et al., 1997; Adams & Wu, 2007) Items auszuschließen, die keine hinreichend gute Passung auf das unterstellte IRT-Modell (hier: Raschmodell) aufweisen. Dieses Fitkriterium ist aber nur dann sinnvoll zu interpretieren, wenn der Test als Ganzes (also die überwiegende Menge der Items) raschhomogen ist. Wenn allerdings sämtliche Items oder ein Großteil der eingesetzten Items durch Kontexteffekte beeinflusst sind (wie es etwa bei Positionseffekten der Fall ist), können Item- und Personenparameter in substanzieller Weise verfälscht sein, ohne dass dies etwa an auffälligen Infitwerten erkennbar wäre. Das Fehlen absoluter Gütemaße für IRT-Modelle (wie etwa der CFI oder TLI für Strukturgleichungsmodelle) erschwert eine Beurteilung, inwieweit das unterstellte IRT-Modell die Daten hinreichend gut beschreibt. Insbesondere die *NAEP reading anomaly* (Beaton, 1988; Zwick, 1991) hat verdeutlicht, dass das einschlägige Instrumentarium an Modellprüfungskontrollen – Infit/Outfit, *differential item functioning* (DIF; Sireci & Rios, 2013), Prüfung auf Linkingfehler (Mazzeo & von Davier, 2008; von Davier, A. et al., 2008) – oft nicht ausreichend ist, um die Güte des Modells bezüglich praktischer Konsequenzen abzuschätzen. Die vorliegende Arbeit konzentriert sich folglich auf die Anwendung von Methoden, die über die einschlägigen Modellprüfungen hinausgehen und es ermöglichen, sowohl das Auftreten von Kontexteffekten zu identifizieren, deren Konsequenzen abzuschätzen und Testdesigns zu evaluieren, die die praktischen Auswirkungen von Kontexteffekten minimieren sollen. Kontexteffekte wurden dabei meist exemplarisch anhand von Positionseffekten be-

trachtet, obwohl viele weitere mögliche Kontexteffekte denkbar sind und auch bereits in empirischen Daten modelliert werden konnten – etwa Testheft- oder Testleteffekte (Frey & Bernhardt, 2012; Hecht et al., 2014), Motivationseffekte (Baumert & Demmrich, 2001; Eklöf, 2010; Wise & DeMars, 2005; Wise & Smith, 2011), Carryover-Effekte (Tuerlinckx & De Boeck, 2004; Yousfi & Böhme, 2012). Viele der Schlussfolgerungen, die hier bezüglich der Positionseffekte gezogen wurden, sind auch auf andere Kontexteffekte übertragbar.

Dass Kontexteffekte in Large-Scale Assessments auftreten, ist eher die Regel als die Ausnahme (Brennan, 1992). Wenn Kontexteffekte mit einer Varianz größer als Null auftreten und durch das Messmodell *nicht* parametrisiert werden, sind Parameter auf der Item- und/oder Personenseite praktisch immer verfälscht. Diese Verfälschung kann auf zwei Ursachen zurückgeführt werden:

1. die Verwendung eines inadäquaten Testdesigns und
2. die Varianzeinschränkung in unterkomplexen IRT-Modellen.

Für die erste Ursache gilt: Sie bewirkt einen deutlich größeren Bias als die zweite Ursache, kann jedoch durch die Verwendung angepasster Testdesigns ausgeschlossen werden. Für die zweite Ursache gilt: Sie tritt praktisch immer auf. Wie groß der Betrag dieses Bias aufgrund der Varianzeinschränkung ist, hängt dabei von der Größe der Varianz der Kontexteffekte ab. In den hier betrachteten Fällen war die Varianz der Itempositionseffekte für Daten des IQB-Ländervergleichs 2011 in der Primarstufe für den Kompetenzbereich Lesen nur sehr klein, und der Bias infolge dessen sehr gering und praktisch nicht bedeutsam.

Da Kontexteffekte sich nicht direkt anhand von Modellgütekriterien identifizieren lassen (vgl. Frey & Bernhardt, 2012), ist es für die Gewährleistung praktisch unverfälschter Item- und Personenparameter wichtig, alle möglichen relevanten Kontexteffekte vorab in konfirmatorischen IRT-Modellen zu identifizieren, um die Fragen zu beantworten: Sind die Kontexteffekte (beispielsweise bezüglich ihrer Effektstärke) so bedeutsam, dass das Testdesign eine Balancierung gewährleisten muss? Und haben die Kontexteffekte eine so große Varianz, dass trotz Balancierung mit einem substanziellen Bias gerechnet werden muss? Die Familie allgemeiner linearer gemischter Modelle (*Generalized Linear Mixed Models, GLMM*) hat sich für solche Modellierungen als nützlich erwiesen, sofern die Kontexteffekte in einparametrischen (1PL) Modellen untersucht werden sollen. Für das Beispiel von Itempositionseffekten zeigte sich dabei für zwei im Rahmen der Evaluation der Bildungsstandards in Deutschland durchgeführte Studien (Pant et al., 2013; Stanat et al., 2012), dass Positionseffekte in einer Weise auftreten, die Balancierung erfordert (vgl. Hecht et al., 2015), verglichen mit den Effekten von Items oder Personen jedoch eine nur geringe Varianz haben. Sofern das Testdesign also Items

und Positionen ausbalanciert, ist der durch Positionseffekte verursachte Bias praktisch unbedeutsam.

Gleichzeitig legen die Ergebnisse des Posterior Predictive Model Checks (Gelman et al., 1996; Rubin, 1984; Sinharay & Haberman, 2014) nahe, dass weitere, hier nicht berücksichtigte Kontexteffekte zusätzlich (oder in Interaktion zu Positionseffekten) auftreten und einen deutlich größeren Bias verursachen könnten. Beispielsweise konnte in dem zweiten Einzelbeitrag eine substanzielle Interaktion von Positionseffekten und der Anstrengungsbereitschaft (*test-taking effort*) nachgewiesen werden. Die Frage der Interaktion von Kontexteffekten und den daraus resultierenden Auswirkungen auf Item- und Personenparameter wurde bislang in der Literatur wenig betrachtet (Debeer & Janssen, 2013). Auch stellen die meisten Simulationsdesigns nur stark vereinfachte Szenarien dar, in denen die wechselseitige Abhängigkeit oder Interdependenz von Kontexteffekten kaum berücksichtigt wird. Wainer und Thissen (1987) bemerken etwa: „[Using] simulated data … tends to trivialize the study, for almost surely the model that is used to generate the data will be the winner in any competition." Gerade die Interaktion von Kontexteffekten auf Item- und Personenseite (*cross-level interaction*) stellt ein gravierendes Problem dar, weil damit eine der zentralen Eigenschaften des Raschmodells aufgehoben wird: Die Unabhängigkeit von Item- und Personenparametern. Über die praktischen Konsequenzen der verletzten Unabhängigkeit können an dieser Stelle keine Aussagen getroffen werden. Die Befunde des zweiten Einzelbeitrags können jedoch genutzt werden, um Simulationsdesigns zu entwickeln, die die Auswirkungen dieser interagierenden Kontexteffekte evaluieren.

Überträgt man die hier angestellten Überlegungen auf die Situation in Large-Scale Assessments, so wird ein Problem deutlich: Um etwa das Testdesign adäquat entwickeln zu können, müsste man bereits vorher wissen, welche Kontexteffekte mit welcher Effektstärke in der Studie auftreten können. Fehler im Testdesign lassen sich später in der Modellierung kaum durch angepasste Auswertungsmodelle korrigieren. Auch ist es nicht ohne Weiteres möglich, einen etwaigen Bias aus Itemparametern des Raschmodells „herauszukorrigieren". Giesbrecht und Gumpertz (2004) betonen: „Although proper examination of the results of an experiment is important, there is no way that a clever analysis can make up for a poorly designed study, a study that leaves out key factors, or inadvertently confounds and/or masks relevant factors." (Giesbrecht & Gumpertz, 2004; S. 2)

Für die empirischen Arbeiten, die etwa am IQB im Rahmen der Evaluation der Bildungsstandards (Granzer et al., 2009; Pant et al., 2013; Stanat et al., 2012; Walther et al., 2007) in verschiedenen Fächern durchgeführt werden, wird damit die Bedeutung des Testdesigns

nochmals deutlich: Aufgrund der gewünschten Anbindung an die Metrik der kriterial bereits definierten Skalen der Kompetenzstufenmodelle ist man auf die Verwendung der diesen Kompetenzmodellen zugrunde gelegten Raschmodelle angewiesen. Im Falle substanzieller Kontexteffekte gibt es nicht die Möglichkeit, alternativ ein Testlet-, Partial Credit- oder Bookletmodell zu verwenden, da die Itemparameter dieser Modelle eine andere Interpretation als im Raschmodell haben und deshalb nicht in Bezug zu den bereits definierten Skalen der Bildungsstandards gesetzt werden können (Tuerlinckx & De Boeck, 2004). Für Analysen im Rahmen der Evaluation der Bildungsstandards erlischt somit die Möglichkeit, Kontexteffekte durch angepasste Messmodelle zu berücksichtigen, hier bleibt nur die Möglichkeit, entsprechend balancierte Testdesigns einzusetzen. Das Testdesign definiert dabei *a priori* eine Gesamtheit von Analysen, die mit den später gewonnenen Daten durchgeführt werden können: Ein Testdesign, das bezüglich der Itemposition nicht balanciert ist, erlaubt keine Analysen von Positionseffekten und dementsprechend keine Aussage darüber, ob dadurch gegebenenfalls mit einem Bias in den Parametern zu rechnen ist. Die Interpretation von Itemparametern aus dem Raschmodell würde dann zwingend voraussetzen, dass keine Positionseffekte existieren. Ein unbalanciertes Testdesign bedeutet also, dass zu den ohnehin restriktiven Annahmen, die der Interpretation der Modellparameter zugrunde liegen, weitere restriktive und *nicht testbare* Annahmen hinzukommen. Adäquate Testdesigns erlauben dagegen nicht nur, die Annahmen zu testen, sondern trotz des Auftretens von Kontexteffekten die unerwünschten Effekte auf die Modellparameter zu minimieren. Ein anderes einfaches Beispiel für eine solche zusätzliche Restriktion wäre ein Versuchsdesign, das die verwendeten Items in zwei Hälften aufteilt, wobei die männlichen Testteilnehmer die eine Hälfte bearbeiten, die weiblichen Testteilnehmer die andere. Die Items wären in den Geschlechtsgruppen genestet (vgl. Kapitel 2.6). Es gäbe nun keine Möglichkeit, Geschlechtsunterschiede in den Testleistungen zu messen; die Interpretation von Item- und Personenparametern in einem solchen Versuchsdesign würde die nicht testbare Annahme implizieren, dass keine Geschlechtsunterschiede existieren. Das Testdesign hätte damit die Anzahl der später möglichen Analysen begrenzt und der Interpretation der Parameter zusätzliche Annahmen hinzugefügt. „Gute“ oder adäquate Testdesigns sollten daher so gestaltet sein, dass sie die Anzahl dieser zusätzlichen, durch sie implizierten und nicht testbaren Annahmen minimieren. Mehr noch: da solche zusätzlichen Annahmen prinzipiell unvermeidbar sind (andernfalls wäre die Menge der benötigten Testhefte unpraktikabel hoch), wäre es wichtig zu evaluieren, welche dieser zusätzlichen Annahmen denn berechtigterweise als gegeben gelten können und nicht durch das Testdesign balanciert und testbar ge-

macht werden müssen. Für Itempositionseffekte gilt jedoch, dass hier eine Balancierung notwendig ist, um unverfälschte Parameter aus dem Raschmodell zu gewinnen.

Die bis hierher diskutierten Punkte sind auch für das Problem fehlender Werte auf den Hintergrundvariablen relevant, das im dritten Einzelbeitrag behandelt wird. Sowohl die Koeffizienten des latenten Hintergrundmodells als auch die manifesten Personenparameter (*plausible values*) werden bedingt auf einen fixierten Vektor von Itemparametern geschätzt. Die notwendige (nicht hinreichende) Voraussetzung für unverfälschte Personenparameter sowie unverfälschte Koeffizienten des latenten Hintergrundmodells sind also unverfälschte Itemparameter. Darüber hinaus können Personenparameter noch durch fehlende Werte auf den Hintergrundvariablen verzerrt werden (Rutkowski, 2011). Ein positiv zu wertender Befund des dritten Einzelbeitrags ist, dass bewährte Verfahren und Methoden zur multiplen Imputation fehlender Werte (Little & Rubin, 1987; Rubin, 1987; van Buuren, 2007) auf Large-Scale Assessments übertragen werden können. Das von Rutkowski (2011) beschriebene Problem eines *tandem shifts* könnte so vermieden werden. In dem dritten Einzelbeitrag wurde der Ausfallprozess dabei jedoch nur linear modelliert. Da die aus Large-Scale Assessments gewonnenen Daten jedoch meist einer Mehrebenenstruktur (*multi level structure*) entsprechen, wäre es denkbar, dass der Ausfallprozess nicht in einem linearen Regressionsmodell, sondern in einem Mehrebenenmodell abgebildet werden müsste. Künftige Forschung könnte sich folglich der Frage widmen, inwieweit die Ausfallprozesse in Large-Scale Assessments tatsächlich über die verschiedenen Ebenen differenziell betrachtet werden müssen.

10. Fazit und Ausblick

In diesem Abschnitt soll kurz auf die in Kapitel 1.2 formulierten Fragen eingegangen werden.

1. Die den statistischen Messmodellen zugrundeliegenden theoretischen Annahmen sind in der Praxis häufig nicht erfüllt: Kontexteffekte treten in Large-Scale Assessments nahezu immer auf, haben jedoch nicht immer praktisch bedeutsame Auswirkungen auf die Güte der Messung.
2. Die potenziell auftretenden unerwünschten Auswirkungen von Kontexteffekten – etwa Verfälschung von Item- und Personenparametern – können mit Hilfe angepasster Messmodelle und/oder angepasster Testdesigns minimiert, nicht jedoch vollständig aufgehoben werden. Da darüber hinaus kein Messmodell und kein Testdesign existiert, das sämtliche möglichen Kontexteffekte adäquat berücksichtigt, ist es einerseits wichtig zu wissen, welche Kontexteffekte in empirischen Anwendungen überhaupt auftre-

ten. Andererseits muss evaluiert werden, welche Kontexteffekte für welche Parameter unter welchen Designbedingungen welche Auswirkungen haben.

3. Die Familie allgemeiner linearer gemischter Modelle (*generalized linear mixed models*; GLMM) eignet sich für die Modellierung verschiedener Kontexteffekte im Rahmen einparametrischer latenter logistischer Messmodelle. Ist ein Kontexteffekt in einem empirischen Anwendungsfall identifiziert, ist damit allerdings noch nichts über dessen praktische Bedeutsamkeit ausgesagt. Zur Beantwortung dieser Frage können Simulationsstudien genutzt werden. Auch eignen sich aus bayesianischen Methoden abgeleitete Verfahren, etwa der *posterior predictive model check* (PPMC), um zu prüfen, inwieweit das parametrisierte Modell zur Replikation von Daten geeignet ist, die den empirisch gewonnenen Daten entsprechen.

In dem ersten Einzelbeitrag dieser Dissertation wurden zwei dieser Punkte – Modellierung von Kontexteffekten und Simulation zur Abschätzung der praktischen Konsequenzen – am Beispiel von Positionseffekten demonstriert. Dabei zeigte sich, dass Positionseffekte in empirischen Studien in einer relevanten Effektstärke auftreten. Die Varianz von Positionseffekten ist dabei jedoch vergleichsweise gering, so dass über das Ausbalancieren von Items und Positionen durch das Testdesign (nahezu) unverfälschte Item- und Personenparameter gewonnen werden können. Diese Befunde sind jedoch nicht ohne weiteres generalisierbar, wenn etwa Positionseffekte mit weiteren Kontexteffekten auf der Item- oder Personenseite interagieren. Dass solche Interaktionen in empirischen Anwendungen auftreten, konnte in dem zweiten Einzelbeitrag gezeigt werden. Das Auftreten von interagierenden Kontexteffekten erschwert Aussagen darüber, wie gravierend ein Bias in Item- oder Personenparametern für einen konkreten Einzelfall ist. In praktischen Anwendungen ist man jedoch häufig gerade daran interessiert. Bayesianische Verfahren wie der PPMC erlauben in solchen Einzelfällen die Einschätzung über das Auftreten eines substanziell bedeutsamen Bias; sie geben jedoch keine Auskunft über dessen Ursache. Im Falle eines solchen Bias wären also weitere konfirmatorische Modelle zu spezifizieren, um Hypothesen über die Ursache des Bias zu testen.

Das Problem fehlender Werte auf Hintergrundvariablen kann in Large-Scale Assessments über Verfahren zur Multiplen Imputation gelöst werden, sofern der Ausfallprozess wenigstens zufällig fehlend (*missing at random*; MAR) und linear oder logistisch linear modellierbar ist, also beispielsweise keiner Mehrebenenstruktur unterliegt. Die Unverfälschtheit der daraus resultierenden Personenparameter und Koeffizienten des latenten Hintergrundmodells hängt jedoch zwingend davon ab, dass die zuvor geschätzten Itemparameter ihrerseits unverfälscht sind. Diese hierarchische Abhängigkeit der Güte der Parameter ist Gegenstand der Kritik (vgl.

Patz & Junker, 1999), jedoch kaum zu vermeiden, wenn für die Analyse frequentistische und nicht bayesianische Verfahren zur Anwendung kommen.

Abschließend soll noch einmal betont werden, dass Kontexteffekte nicht zwangsläufig als ein „Problem“ bei der Leistungsmessung in quantitativer empirischer Bildungsforschung betrachtet werden müssen. Sie können problematisch sein, da man in der Praxis von Large-Scale Assessments häufig die Implikationen der starken IRT-Modellannahmen (*strong assumptions*) für die Konstruktion von Skalen und/oder deren Verlinkung nutzt, obschon diese Annahmen häufig nicht hinreichend gegeben sind. Aus messtheoretischer Perspektive ist der umgekehrte Weg, nämlich diese Annahmen zu modellieren und also zu *testen*, anstatt ihre Implikationen bereits vorauszusetzen, häufig vielversprechender. Die Analyse von Kontexteffekten erlaubt im Idealfall, die den Testaufgaben zugrunde liegenden inhaltlichen Modellvorstellungen kritisch zu überprüfen und zu hinterfragen. Inwieweit es gelungen ist, ein theoretisch definiertes Konstrukt über Testaufgaben zu operationalisieren, kann durch die Analyse von Kontexteffekten beantwortet werden.

Die Verwendung von IRT-Modellen im Rahmen der Evaluation der Bildungsstandards in Deutschland hat den Fokus jedoch hauptsächlich auf die Konstruktion kriterialer Skalen im Rahmen eines *Standard Settings* gelegt, das die Gültigkeit der Modellannahmen bereits zur Voraussetzung hat. Frühere Publikationen zu probabilistischen Messmodellen (vgl. Hambleton et al., 1991) betonen jedoch gerade die Überprüfbarkeit dieser Annahmen als wesentlichen Vorteil der IRT.

Insgesamt unterstreicht die Arbeit die Notwendigkeit der Modellierung von Kontexteffekten. Eine Modellprüfung über einschlägige Gütekriterien (Infit/Outfit, DIF, Q3-Statistik) ist oft nicht hinreichend, um verschiedene relevante Kontexteffekte zu identifizieren. Zum zweiten wird die Anwendung komplexer Simulationsdesigns empfohlen, um die Auswirkung interagierender Kontexteffekte abzuschätzen.

11. Literatur

Die hier verzeichneten Literaturangaben betreffen lediglich die in der Rahmung der vorliegenden Arbeit zitierten Quellen. Die in den Einzelbeiträgen zitierten Quellen werden in jeweils separaten Literaturverzeichnissen im Anschluss an die Einzelbeiträge aufgeführt.

Adams, R. J., Wilson, M. & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Adams, R. J. & Wu, M. L. (2007). The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and Mixture Distribution Rasch Models* (S. 57-75). New York: Springer.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 22*, 47-76.

Allen, N. L., Donoghue, J. R. & Schoeps, T. L. (2001). *The NAEP 1998 Technical Report*. Washington: National Center for Educational Statistics.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research, 28*(2), 186-208.

Asseburg, R. & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92-104.

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441-462.

Beaton, A. E. (1988). *The NAEP 1985-86 Reading Anomaly: A Technical Report*: Educational Testing Service, National Assessment of Educational Progress, Princeton.

Böhme, K. & Robitzsch, A. (2009). Methodische Aspekte der Erfassung der Lesekompetenz. In D. Granzer, O. Köller, A. Bremerich-Vos, M. Van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 250-289). Weinheim: Beltz.

Bonsen, M., Lintorf, K., Bos, W. & Frey, K. A. (2008). TIMSS 2007 Grundschule - Eine Einführung in die Studie. In W. Bos, M. Bonsen, J. Baumert, M. Prenzel, C. Selter & G. Walther (Hrsg.), *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 19–48). Münster: Waxmann.

Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York, NY: Wiley.

Brennan, R. L. (1992). The Context of Context Effects. *Applied Measurement in Education, 5*(3), 225-264.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Hrsg.), *Handbook of research on teaching*. Chicago, IL: Rand McNally.

Cizek, G. J. (2001). Conjectures on the rise and fall of standard setting: An introduction to

context and practice. In G. J. Cizek (Hrsg.), *Setting performance standards: Concepts, methods, and perspectives* (S. 3–17). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications Ltd.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F.et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1-28.

De Boeck, P. & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory Item Response Models* (S. 3-42). New York: Springer.

Debeer, D. & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185.

Duncan, T. E., Duncan, S. C. & Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling*. Mahwah: Erlbaum.

Dziak, J. J., Coffman, D. L., Lanza, S. T. & Li, R. (2012). *Sensitivity and Specifcity of Information Criteria*: The Pennsylvania State University.

Eklöf, H. (2010). Student Motivation and Effort in the Swedish TIMSS Advanced Field Study, *IEA International Research Conference*.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359-374.

Fischer, G. (1974). *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen*. Bern: Huber.

Fox, J.-P. (2010). *Bayesian Item Response Modeling. Theory and Applications*. New York, Dordrecht, Heidelberg, London: Springer.

Fox, J.-P. & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika, 66*(2), 271-288.

Foy, P., Galia, J. & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science asssessment. In J. F. Olson, M. O. Martin & I. V. S. Mullis (Hrsg.), *TIMSS 2007 Technical Report* (S. 225-280). Chestnut Hill, MA: TIMSS & PIRLS

International Study Center, Lynch School of Education, Boston College.

Frey, A. & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling, 54*(4), 397-417.

Frey, A., Carstensen, C. H., Walter, O., Rönnebeck, S. & Gomolka, J. (2008). Methodische Grundlagen des Ländervergleichs. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2006 in Deutschland: Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (S. 375–397). Münster: Waxmann.

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53.

Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-807.

Giesbrecht, F. G. & Gumpertz, M. L. (2004). *Planning, Construction, and Statistical Analysis of Comparative Experiments*. Hoboken, N.J.: Wiley.

Glas, C. A. W. & Suarez Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*(2), 87-106.

Gonzalez, E. & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph, 3*, 125-156.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Granzer, D., Köller, O., Bremerich-Vos, A., Van den Heuvel-Panhuizen, M., Reiss, K. & Walther, G. (Hrsg.). (2009). *Bildungsstandards Deutsch und Mathematik*. Weinheim und Basel: Beltz.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology, 4*, 75-89.

Harel, O. & Schafer, J. L. (2003). Multiple imputation in two stages,*Proceedings of Federal Committee on Statistical Methodology Research Conference* (S. 91-96). Washington DC.

Hartig, J. & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54*(4), 418-431.

Hecht, M., Weirich, S., Siegle, T. & Frey, A. (2014). Modeling booklet effects for nonequivalent group designs in large-scale assessment. *Educational and Psychological Measurement*, 1-17.

Hecht, M., Weirich, S., Siegle, T. & Frey, A. (2015). Effects of Design Properties on Parameter Estimation in Large-Scale Assessments. *Educational and Psychological Measurement*, 1-24.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L. & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*(3), 391-402.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Jiao, H., Wang, S. & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50*(2), 186-203.

Khorramdel, L. & Frebort, M. (2011). Context effects on test performance: What about test order? *European Journal of Psychological Assessment, 27*(2), 103-110.

Kingston, N. M. & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*(2), 312-320.

Kolen, M. J. & Brennan, R. L. (2004). *Testing equating, scaling, and linking: Methods and practice*. New York: Springer.

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Burlington, MA: Academic Press.

Leary, L. F. & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*(3), 387-413.

Lin, T. H. & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*(3), 249-264.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association, 87*(420), 1227-1237.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analyses with Missing Data*. New York: Wiley.

Lord, F. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA:

Addison-Wesley.

Marsh, H. W. (1984). Experimental manipulations of university student motivation and their effects on examination performance. *British Journal of Educational Psychology, 54*(2), 206-213.

Mazzeo, J. & von Davier, M. (2008). Review of the Programme for International Student Assessment (PISA) Test Design: Recommendations for fostering stability in assessment results.: Available at: doc.ref. EDU/PISA/GB(2008)28.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99-114.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741-749.

Messick, S. (1998). Test validity: A matter of consequence. In B. D. Zumbo (Hrsg.), *Validity Theory and the Methods Used in Validation: Perspectives From the Social and Behavioral Sciences* (S. 35-44). The Netherlands: Kluwer Academic Press.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.

Molenberghs, G. & Verbeke, G. (2004). An Introduction to Generalized (Non)Linear Mixed Models. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory Item Response Models* (S. 111-166). New York: Springer.

Monseur, C., Baye, A., Lafontaine, D. & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, 4*, 131-155.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*(1), 67-82.

Muthén, L. K. & Muthén, B., O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133-142.

OECD. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris, France: OECD.

OECD. (2012). *PISA 2009 Technical Report.* Verfügbar unter: http://dx.doi.org/10.1787/9789264167872-en [18.07.2012

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (Hrsg.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster, New York, München, Berlin: Waxmann.

Patz, R. J. & Junker, B. W. (1999). A straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

Penfield, R. D. & Camilli, G. (2007). Differential Item Functioning and Item Bias. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics* (S. 125-168). New York: Elsevier.

Penk, C., Pöhlmann, C. & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific differences. *Large-scale Assessments in Education, 2*(5), 2-17.

Perlini, A. H. & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology, 39*(4), 299-307.

Plummer, M. (2013). JAGS - Just another Gibbs sampler (Version 3.4.0).

Plummer, M. (2014). rjags: Bayesian graphical models using MCMC. R package version 3.14.

R Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.0). Vienna, Austria: R Foundation for Statistical Computing.

Reiter, J. P. & Drechsler, J. (2007). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *IAB discussion paper, 2007*(20).

Reiter, J. P. & Raghunathan, T. E. (2007). The multiple Adaptions of Multiple Imputation. *Journal of the American Statistical Association, 102*, 1462-1471.

Richter, D., Engelbert, M., Böhme, K., Haag, N., Hannighofer, J., Reimers, H.et al. (2012). Anlage und Durchführung des Ländervergleichs. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 85-102). Münster: Waxmann.

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* (2.). Bern: Huber.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistican. *The Annals of Statistics, 12*(4), 1151-1172.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica, 57*(1), 3-18.

Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement, 48*(3), 293-312.

Ryan, K. E. & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*(1), 73-90.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Schroeders, U., Robitzsch, A. & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-Tests. *Journal of Educational Measurement, 51*(4), 400-418.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-465.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement, 42*(4), 375-394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*, 429-449.

Sinharay, S. & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*(1), 23-35.

Sinharay, S. & Johnson, M. S. (2003). *Assessing Fit of Unidimensional Item Response Theory Models Using a Bayesian Approach*. Princeton, NJ: Educational Testing Service.

Sireci, S. G. & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2-3), 170-187.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B - Statistical Methodology, 64,* 583-616.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik*. Münster: Waxmann.

Stewart, E. E. (1981). Methodological issues related to the study of context effects in multisection tests,*annual meeting of the National Council on Measurement in Education*. Los Angeles, CA.

Stoel, R. D., Galindo Garre, F., Dolan, C. & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11*(4), 439-455.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to undimensionality assessment and ability estimation. *Psychometrika, 55*(2), 293-325.

Stout, W., Nandakumar, R., Junker, B. W., Chang, H. H. & Steidinger, D. (1991). DIMTEST: A Fortran Program for assessing Dimensionality of Binary Item Responses: University of Illinois, Department of Statistics, Champaign.

Swaminathan, H., Hambleton, R. K. & Rogers, H. J. (2007). Assessing the Fit of Item Response Theory Models. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics* (Bd. 26, S. 683-718).

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*(3), 159-203.

Tuerlinckx, F. & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory Item Response Models* (S. 289-316). New York: Springer.

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noortgate, W.et al. (2004). Estimation and software. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory Item Response Models* (S. 343-373). New York: Springer.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*, 219-242.

Van der Linden, W. J., Veldkamp, B. P. & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28*(5), 317-331.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York, NY: Springer.

von Davier, A., Carstensen, C. H. & von Davier, M. (2008). Linking Competencies in Horizontal, Vertical, and Longitudinal Settings and Measuring Growth. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 121-149). New York: Hogrefe & Huber.

von Davier, M., Gonzalez, E. & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large Scale*

Assessments, 2, 9-36.

Von Davier, M., Sinharay, S., Oranje, A. & Beaton, A. E. (2007). The Statistical Procedures Used in National Assessment of Educational Progress: Recent Developments and Future Directions. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics* (Bd. 26, S. 1039-1056). Amsterdam: Elsevier.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for two testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H. & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*(4), 339-368.

Walther, G., Van den Heuvel-Panhuizen, M., Granzer, D. & Köller, O. (2007). *Bildungsstandards für die Grundschule: Mathematik konkret*. Berlin: Cornelsen.

Weirich, S., Haag, N. & Roppelt, A. (2012). Testdesign und Auswertung des Ländervergleichs: Technische Grundlagen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 277-290). Münster: Waxmann.

Weirich, S. & Pant, H. A. (2014). *Vergleichsarbeiten 2014, 3. Jahrgangsstufe Deutsch. Technischer Bericht*: Institut zur Qualitätsentwicklung im Bildungswesen.

Wilson, M. & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory Item Response Models* (S. 43-74). New York: Springer.

Wise, S. L. & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assesment, 10*(1), 1-17.

Wise, S. L. & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

Wise, S. L. & Smith, L. F. (2011). A Model of Examinee test-Taking Effort. In J. A. Bovaird, K. F. Geisinger & C. W. Buckendahl (Hrsg.), *High-stakes testing in education: Science and practice in K-12 settings* (S. 139-153). Washington, DC: American Psychological Association.

Woods, C. M. & Harpole, J. (2014). How item residual heterogeneity affects tests for differential item functioning. *Applied Psychological Measurement*, 1-13.

Wu, M., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER ConQuest 2.0 - Generalized item response modelling software*. Camberwell: ACER.

Yen, W. M. (1980). The Extent, Causes and Importance of Context Effects on Item Parameters

for Two Latent Trait Models. *Journal of Educational Measurement, 17*(4), 297-311.

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8*(2), 125-145.

Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Yousfi, S. & Böhme, H. F. (2012). Principles and procedures of considering item sequence effects in the development of calibrated item pools: Conceptual analysis and empirical illustration. *Psychological Test and Assessment Modeling, 54*(4), 366-393.

Zumbo, B. D. (2007). Validity: Foundational Issues and Statistical Methodology. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics* (Bd. 26, S. 45-80).

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(10-16).

A. Anhang A, Beitrag 1: Modeling Item Position Effects Using Generalized Linear Mixed Models

Dieser Beitrag ist in der Zeitschrift *Applied Psychological Measurement* erschienen. Die Referenz lautet:

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*(7), 535-548. doi: 10.1177/0146621614534955

Der Link für den Download des Beitrags ist:
http://apm.sagepub.com/content/38/7/535

B. Anhang B, Beitrag 2: Item Position Effects are Moderated by Changes in Test-Taking Effort[3]

Item Position Effects are Moderated by Changes in Test-Taking Effort

Sebastian Weirich, Christiane Penk, Martin Hecht, Alexander Roppelt & Katrin Böhme

Author Note

Sebastian Weirich, Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin; Christiane Penk, Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin; Martin Hecht, Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin; Alexander Roppelt, Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin; Katrin Böhme, Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin

Correspondence concerning this article should be addressed to Sebastian Weirich, Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, Email: sebastian.weirich@iqb.hu-berlin.de

[3] Diese Artikelfassung wurde im November 2014 erstellt und ist in dieser Version noch nicht zur Publikation angenommen worden. Dies ist nicht die Originalversion des Artikels. Diese Fassung kann nicht zur Zitierung herangezogen werden.

Abstract

This article examines the interdependency of two context effects which are known to regularly occur in large-scale assessments: item position effects and effects of test-taking effort on the probability of correctly answering an item. To date, these effects have only been considered separately. We used a micro-longitudinal design to measure the initial effort and the change in effort at three points in time during the course of a large-scale assessment test of 120 minutes. We found that the current test-taking effort diminishes considerably during the course of the test. Moreover, we found substantial non-linear position effects which indicate that the item difficulty increases during the test. Furthermore, persons' competence estimates differ across positions. This fluctuation is associated both with initial effort and change in effort: Position effects are more pronounced for persons with lower initial effort and for persons whose test-taking effort declines faster. Consequences of these results concerning reliability and validity of large-scale assessments are discussed.

Keywords: item position effects, generalized linear mixed models, test-taking effort

Recently, several methodological studies examined item position effects in an item response theory (IRT) framework (Albano, 2013; Debeer & Janssen, 2013; Hahne, 2008; Hohensinn et al., 2008; Hohensinn, Kubinger, Reif, Schleicher, & Khorramdel, 2011; Meyers, Miller, & Way, 2009; Schweizer, Schreiner, & Gold, 2009; Weirich, Hecht, & Böhme, 2014). Item position effects, which belong to the group of context effects (Brennan, 1992; Wainer & Kiely, 1987), imply that the difficulty of an item in an achievement test varies depending on of the position of the item in the booklet. Usually, an item administered at the end of a test is more difficult than the same item administered at the beginning of the test. If an achievement test—for example in large-scale assessments (Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010) or in computerized adaptive testing (CAT; Wainer & Kiely, 1987)—consists of several test forms which may include different items in a different order and which are to be linked to a common scale, item position effects might violate the assumption of item parameter invariance across test forms.

As the assumption of item parameter invariance is central in any linking procedure based on common items (Cook & Petersen, 1987; Kolen & Brennan, 2004) item position effects might induce linking bias (Debeer & Janssen, 2013; Meyers et al., 2009). To date, the focus of research concerning item position effects lies on three topics: first, to examine how severely item parameter estimates may be biased (Debeer & Janssen, 2013; Meyers et al., 2009), second, to develop and to evaluate balanced test designs which are proposed to minimize this bias (Frey et al., 2009; Gonzalez & Rutkowski, 2010; Weirich et al., 2014), and third, to develop models which are suited to estimate effects of item position in an IRT framework (De Boeck et al., 2011; Debeer & Janssen, 2013; Janssen, Schepers, & Peres, 2004; Tuerlinckx & De Boeck, 2004).

In the following, central results regarding these three topics will be briefly reviewed. Meyers et al. (2009) compared item parameters of two tests where the overlapping items differ in their relative position. They found that about 56% of the variance in change in Rasch item difficulty could be attributed to item position change. For example, items whose positioning differed more than 20 positions between both test, changed about 0.20 logits on average. Debeer and Janssen (2013) used data from the Pisa 2006 assessment (OECD, 2006) to show that the difficulty of reading items increased .24 logits on average when administered one cluster position further in the test. Moreover, their simulation yielded that position effects may cause a bias in item parameter estimates of the Rasch model of up to .52 logits.

Concerning the second topic, research on test designs yielded that this bias can be minimized by the application of test designs which are balanced with respect to item position

(Weirich et al., 2014). If several test forms (e.g., booklets) are constructed and distributed to examinees in such a manner that each item occurs at each position an equal number of times, the disturbing effect of item position is expected to influence each item in the same magnitude. Hence, there is no differential effect of item position on item difficulty. However, in spite of balancing, an effect of item position on item difficulty on average cannot be ruled out. The third topic addresses the development and evaluation of models to estimate item position effects. The framework of generalized linear mixed models (GLMM; De Boeck & Wilson, 2004) has proved to be very useful in modeling item position effects when only 1PL estimation is of interest (Debeer & Janssen, 2013; Weirich et al., 2014). The position of an item may be specified as a predictor to examine whether the probability of success depends on it. Position effects may be specified to be linear or nonlinear. Moreover, if items and persons are considered as random effects, the distribution of items or persons can be specified to be dependent on position, for example to investigate whether the effect of the position is homogeneous or heterogeneous across persons or items.

When asking why item position effects occur, Debeer and Janssen (2013) describe a lack of research. Depending on the direction of the effect two possible explanations are usually taken into account (Hohensinn et al., 2011; Kingston & Dorans, 1984): An increase of item difficulty during the test might be interpreted as a fatigue effect or an effect of decreasing test-taking effort (for example, due to decreasing motivation) of the examinees. On the other hand, a decrease of item difficulty may be interpreted as a practice effect, for example if examinees become more acquainted with the test material. Although these explanations seem plausible at first glance, Debeer and Janssen (2013) suggest to answer “the why question” by including person predictor variables, for example the test-taking effort in the model. Such a model then resembles a differential item functioning (DIF) model which allows for tests as to whether effects of items are invariant across subpopulations. Correspondingly, a ”differential position functioning“ model could investigate whether effects of item positions are invariant across subpopulations (for example, examinees who invest more vs. less effort). However, item position effects conceptually imply a change in the item response behavior during the course of the test: Examinees increasingly reduce test-taking effort, or they progressively develop test wiseness that allows them to improve their performance. Consequently, as item position effects are ultimately an effect of test time, it is plausible to assume that covariates which substantively moderate these effects also change over time. Following Debeer and Janssen (2013), the change in effort of examinees during the test is very plausible to also moderate item position effects.

Like position effects, effects of the examinees' effort on the test score may be considered as another context effect. A relation between self-reported effort and test achievement was found in several studies (for an overview, see Wise & DeMars, 2005). In general, students who report higher motivation invest more effort and achieve higher scores in the test. The topic is frequently discussed as a problem of validity: If achievement tests are low-stakes to the examinees (i.e., test results have no consequences for examinees as they neither receive grades nor academic credits), it cannot be guaranteed that they will do their best. Hence, a lack of test-taking effort not only leads to a potential underestimation of achievement scores but also to a change of the construct being measured. The test score then would be comprised of two constructs: the ability and the current effort of the examinees (Eklöf, 2010), i.e. test scores would not be free from construct-irrelevant variance (Messick, 1984). When interpreting the relation between self-reported motivation and test achievement, it is unclear whether highly motivated students perform better due to their higher motivation or due to their higher actual ability (Wise & DeMars, 2005). Measuring the change in test-taking effort during the course of the test in a micro longitudinal design is an approximation to an experimental design and allows examining as to whether situational motivation affects the current response behavior beyond the general correlation between motivation and ability.

To summarize: Position effects imply a change in the response behavior. If these effects are moderated by a change in test-taking effort, we could reason an effect of effort on test performance beyond the examinees' ability. Position effects and effort both are a possible source of construct-irrelevant variance in test scores. A better understanding of the potential interplay of both effects can help to avoid biased person estimates in large-scale assessments.

To distinguish between item test-taking motivation and test-taking effort, we adopt the taxonomy of Freund et al. (2011). Test-taking motivation is considered as a multidimensional construct which also comprises facets as test related anxiety, challenge, interest and perceived probability of success. Test-taking effort, as a facet of test-taking motivation, is a unidimensional state which may change during the course of the test (Wise & Smith, 2011).

In the present article, we use data from the German National Assessment Study, a German nation-wide low-stakes large-scale assessment study similar to the US National Assessment of Educational Progress (NAEP). The tasks and items included in this study were developed to evaluate student attainment according to the German National Educational Standards. Test-taking effort was measured three times during the test. The change in test-taking effort was modeled using a latent growth model assuming a linear trend during the course of the test. Intercept and slope of the growth model are used as estimates for initial effort and change in ef-

fort. We model position effects in a generalized linear mixed effects model framework and examine whether position effects are moderated by initial effort and change in effort. Hence, we investigate the interaction between position effects and test-taking effort, which have only been considered separately, so far.

We address are the following research questions:

1. To which degree does the test-taking effort of examinees change during a low-stakes achievement test of 120 minutes?
2. Do position effects occur in a 120-minutes low-stakes achievement test? Are examinees affected heterogeneously by position effects?
3. Are position effects moderated by initial effort and change in effort?

Method

Sample: We used data collected for the German National Assessment Study 2012 (Pant et al., 2013). In a multistage sampling procedure, schools were randomly selected within each of the 16 German federal states; within each school, one or two Grade 9 classes were randomly selected. As several domains in several school types were tested, the entire design was composed of several subdesigns (Hecht, Roppelt, & Siegle, 2013). To reduce the amount of data and to simplify the analyses, only data from one domain (scientific literacy) and from several school types (secondary school, comprehensive school) were used. Our subsample of N=9,410 9th graders (51.4% female, mean age=15.7 years) consists of examinees who answered the effort scale at all three points of time.[4]

Measures: In the test 386 dichotomously scored scientific literacy items were used. In a preceding multistage development process, items were developed and selected to fit the unidimensional Rasch model. Unsuitable items were discarded, so the items used in this study have proved to fulfill the measurement standards of the Rasch model. The scientific literacy items were used in a multiple matrix sampling design in which a subset of items (i.e., a booklet) was randomly assigned to each examinee. More specifically, the items were grouped into 31 disjoint blocks, i.e. items were nested in blocks. The time allocated for each block was 20

[4] In the German National Assessment Study 2012 test-taking effort of $N = 47{,}739$ examinees overall was measured. Due to a multiple matrix sampling design of questionnaires only a subset of examinees answered the effort scale at all three points of time. Moreover, we excluded 1,208 examinees who used the same option for all items of all motivation scales (no matter whether the item had a positive or negative formulation). These examinees were suspected to have had responded carelessly or untruthfully and were excluded from the analysis. Overall, $N = 20{,}934$ examinees with valid answers at all three points of time remained. Only half of them was provided with test booklets of science tasks, the other half worked on math tasks. Hence, $N = 9{,}410$ examinees who had worked on science tasks and completed motivation items at all three points of time, remained for the main analysis.

minutes. 31 booklets were constructed, each consisting of 6 blocks, according to a Youden square design (Frey et al., 2009), which is balanced with respect to item position. Hence, the time allocated for each booklet was $6 \times 20 = 120$ minutes, where a 15 minutes intermission was scheduled after 60 minutes to provide examinees with the opportunity to relax. As common in such designs, only the position of blocks is varied across booklets, whereas the position of items within each block remains constant. Modeling position effects therefore means to estimate the magnitude by which item difficulty changes on average if items are placed, for example, at block position 2 instead of block position 1. The 31 booklets were randomly distributed among the 9,410 students, yielding 692,595 responses overall.

Test-taking effort was measured in a self-report, using the test-taking effort scale examined in Penk, Pöhlmann, and Roppelt (2014), who investigated test-taking motivation as a multidimensional construct. Test-taking effort constitutes one facet of this construct and was measured with four items which originally stem from Eklöf (2010). All items ranged from 1 = strongly disagree to 4 = strongly agree and referred to the current test situation. All items were adopted (positive formulation only) and translated into German (Penk et al., 2014).

Test-taking effort was measured three times during the test, at the beginning (i.e. prior to examinee's work on the test; t1), after the examinees had finished the half of the test booklet (i.e. after 60 minutes of test processing; t2), and after examinees had finished the whole test (t3). Scale reliability of the four effort items yielded satisfying results. We found reliability coefficients of Cronbach's Alpha (Cronbach, 1951) of .82 for t1, .86 for t2, and .83 for t3.

Models: First, the change in test-taking effort was modeled using a latent linear growth model in Mplus, version 7.11 (Muthén & Muthén, 1998-2012). We specified a curves-of-factors model (Duncan, Duncan, & Strycker, 2006), which assumes equal intercepts and equal factor loading across the three points of time. This corresponds to the assumption of strong invariance. We tested whether this assumption is supported by the data using the Wald test.

Afterwards, we drew factor scores for both parameters estimated in the model, the intercept (initial effort) and the slope of the linear change in effort for each examinee. The factor score variables subsequently were used as covariates in the GLMMs. Both factor score variables were standardized in order to simplify the interpretation of parameter estimates in the following GLMMs.

Second, we used the R package *lme4* (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014) to specify three nested GLMMs to model nonlinear item position effects dependent on initial effort and change in effort. In each subsequent model, additional predictors are included, whereas the random effect structure remains unchanged. Model 1 only assumes

nonlinear effects of item position. Model 2 additionally uses the initial effort and the change in effort from the latent linear growth model as predictors. In model 3, two-way interactions (e.g., cross-level interactions) of position and effort are parametrized. Hereafter, the three models are described in more detail. We will derive them from the simple Rasch model, which consists of three components (see, for example, De Boeck et al., 2011):

Random component: A specific binary item response X_{ji} for person j and item i is generated from a binomial distribution with one observation and a probability π_{ji}:

$$X_{ji} \sim \text{binomial}(1, \pi_{ji}),$$

Linking component: π_{ji} is the probability that person j solves item i, hence: $\pi_{ji} = \left(P(X_{ji} = 1)\right)$. To map the interval [0, 1] of π_{ji} into $[-\infty, +\infty]$, the logit transformation is used:

$$\eta_{ji} = \text{logit}(\pi_{ji}),$$

Linear component: The transformed probability η_{ji} now is the dependent variable in the Rasch model:

$$\eta_{ji} = \alpha_0 + \theta_j - \beta_i,$$

for persons $j = 1, \ldots, J$, and for items $i = 1, \ldots, I$. The person's ability is denoted as θ_j, and the item difficulty is denoted as β_i. η_{ji} is the logit for the probability that person j correctly solves item i.

According to De Boeck (2008), we may distinguish four different kinds of Rasch models, depending on whether the persons and the items are either treated as random or fixed. When using the Rasch model as a measurement model, a random person—fixed item specification is appropriate where $\alpha_0 = 0$. For each of the I items, a difficulty parameter may then be estimated, then. The persons are considered as random, assuming they are randomly sampled from a normally distributed population: $\theta_j \sim N(0, \sigma_\theta^2)$. In our context, however, a random person—random item specification (De Boeck, 2008) is appropriate, as we are less interested in effects of single items than in effects of single positions. Hence, we assumed that the items are sampled from a pool of items with normally distributed difficulty as well: $\beta_i \sim N(0, \sigma_\beta^2)$. As the mean of the random effects θ_j and β_i both equals zero, the intercept term α_0 describes the difference between mean item difficulty and mean person ability. To estimate position effects, we extended the random person—random item Rasch model by a further component, Δ:

$$\eta_{ji} = \Delta + \theta_{jp} + \beta_i, \tag{B1}$$

for positions $p = 1, \ldots, P$. We used the "+ - parametrization" which leads to an interpretation of β_i as the easiness of an item. We were interested in the effects of specific positions, so we modeled the corresponding effect of Δ to be fixed. Moreover, as we were interested in modeling non-linear position effects, the variable p which indicates the item's position (with $p = 1, \ldots, P$) was decomposed into P dichotomous indicator variables. These form the indicator matrix A with dimensions $N \times P$, where P equals the number of positions, and N equals the total number of all responses to all items. The first column of A (A_1) equals 1 for each response. For $p \geq 2$ the p^{th} column of A (A_p) equals 1 for all those responses which occurred at position p, 0 otherwise. To derive our first model, we may formulate the position effect for the n^{th} response as follows (De Boeck et al., 2011; De Boeck & Wilson, 2004):

$$\Delta = \sum_{p=1}^{P} \delta_p A_{(j,i)p} \ . \tag{B2}$$

δ_1 now can be interpreted as the intercept which was formerly denoted α_0. δ_2 can be interpreted as the effect of position 2 versus position 1. Substituting Equation B2 in Equation B1 yields our first model. θ_j and β_i may be specified as unconditional random effects, i.e. $\beta_i \sim N\left(0, \sigma_\beta^2\right)$. However, as we were interested in a potential interdependency of position effects and effort (i.e., a person-specific variable), it was plausible to assume the person effect θ_j to be conditional on position to answer the question: Does the distribution of persons vary across positions? In other words, we allowed for multidimensionality on the person side. θ_{jp} therefore was assumed to be jointly multivariate normally distributed. We used the "p" as a second subscript for θ to emphasize that θ was heteroscedastic across positions:

$$\theta_{jp} \sim MVN\left(0, \Sigma\right), \quad \text{where } \Sigma = \begin{pmatrix} \sigma_{11} & & \\ \vdots & \ddots & \\ \sigma_{P1} & \cdots & \sigma_{PP} \end{pmatrix}. \tag{B3}$$

In *lme4*, model 1 was specified as

```
value ~ position + (-1 + position | subject) + (1|item).
```

Additionally, Model 2 uses the factor scores of the initial effort I_j and the change in effort C_j from the latent linear growth model as fixed-effects predictors. Two additional model parameters, ρ and φ will be needed.

$$\eta_{jip} = \Delta + \rho I_j + \varphi C_j + \theta_{jp} + \beta_i, \tag{B4}$$

with $\Delta = \sum_{p=1}^{P} \delta_p A_{(j,i)p}$. We used the subscript j for I and C as both variables are person variables. Note that Roman—instead of Greek—letters are used to emphasize the distinction be-

tween estimated model parameters (Greek letters) and manifest variables (Roman letters). The random effects structure was the same as in model 1. In *lme4*, model 2 was specified as

```
value ~ position + initial + change + (-1 + position | subject) + (1|item).
```

In model 3, we additionally assumed two-way interactions between position effects and both intercept and slope terms. For this purpose, we parametrized

$$\Delta = \sum_{p=1}^{P} \left(\delta_p A_{(j,i)p} + \omega_p A_{(j,i)p} I_j + \psi_p A_{(j,i)p} C_j \right), \tag{B5}$$

where ω_p and ψ_p are the parameters for the interaction between the effect of position p and the initial effort I_j and the change in effort C_j, respectively. Substituting Equation B5 in Equation B4 yields model 3. Again, the random effects structure remains unchanged. In *lme4*, model 3 was specified as

```
value ~ position + position:initial + position:change + initial + change
        + (-1 + position | subject) + (1|item).
```

Results

Table B1 and Table B2 list the results which refer to the first research question. Table B1 lists the factor loadings both with and without the assumption of strong invariance. Assuming equal intercepts and equal factor loadings for the indicators across the three points of time—which corresponds to the assumption of strong invariance—is not strictly feasible. We tested the restriction against a more liberal model which assumes different factor loadings for each of the four items at each point of time. The Wald test yields a significant loss of model fit due to the restriction: $\chi^2 = 425.67$; $df = 6$; $p < .001$. However, the factor loadings of both models differ only slightly. Moreover, the difference between the slope coefficients is negligible (standardized: –0.76 vs. –0.78). Hence, we decided to maintain the model assumption of strong invariance because we are interested in interpreting score changes across points of time due to changes in effort.

Table B2 displays the coefficients of the linear growth model. On average, test-taking effort diminishes during the course of the test. The standardized coefficient of the slope is –0.76, which indicates a moderate decline in the examinees' effort. Moreover, we see that the slope variance is quite low compared to the intercept variance: Whereas the base level effort is rather heterogeneous between examinees, the decrease in effort is more homogeneous between examinees. Descriptive results of the factor scores yielded that only 12.1 percent of the total sample showed an increase in test-taking effort during the course of the test.

Table B1: Fixed and free factor loadings for the latent linear growth model of change in test-taking motivation

Parameter	**invariance assumed**	**invariance not assumed**		
Factor loading	Est.	Est.t1	Est.t2	Est.t3
E1	1.000	1.000	1.000	1.000
E2	0.719	0.665	0.812	0.688
E3	1.046	1.049	1.121	1.006
E6	1.097	1.117	1.182	1.046

Note. E1, E2, E3 and E6 refer to the notation of items measuring test-taking effort as reported in Table 2 of Eklöf (2010)

Table B2: Results of latent linear growth model of change in test-taking motivation

Parameter					
	Unstandardized coefficients			Standardized coefficients	
Effect	Est.	SE	*p*	Est.	SE
Intercept	0.000				
Slope	-0.173	0.003	< .001	-0.764	0.015
Var(Intercept)	0.404	0.007	< .001	1.000	
Var(Slope)	0.051	0.001	< .001	1.000	
cor(Intercept and Slope)	-0.109	0.013	< .001	-0.109	0.013
CFI		0.947			
TLI		0.944			

The results of the three GLMMs are listed in Table B3. In model 1, only a nonlinear item position effect was specified, which refers to the second research question. The regression coefficients of δ_2 up to δ_6 represent the mean change in the logit for a correct response of each position in relation to the first position. Negative coefficients indicate that the logit for a correct response decreases with increasing position. For example, $\delta_3 = -0.37$ means that the logit for a correct response at position 3 is on average 0.37 lower than the logit at position 1. Note

Table B3: Fixed and random effects for the three GLMMs

Parameter	**Model 1**			**Model 2**			**Model 3**		
Fixed effects	Est.	SE	*p*	Est.	SE	*p*	Est.	SE	*p*
δ_1 (Intercept)	0.176	0.064	0.006	0.177	0.064	0.006	0.176	0.064	0.006
δ_2: Position 2	-0.096	0.011	< .001	-0.097	0.011	< .001	-0.096	0.011	< .001
δ_3: Position 3	-0.372	0.011	< .001	-0.373	0.011	< .001	-0.371	0.011	< .001
δ_4: Position 4	-0.161	0.011	< .001	-0.162	0.011	< .001	-0.161	0.011	< .001
δ_5: Position 5	-0.251	0.011	< .001	-0.252	0.011	< .001	-0.250	0.011	< .001
δ_6: Position 6	-0.383	0.011	< .001	-0.384	0.011	< .001	-0.382	0.011	< .001
ρ : initial effort				0.299	0.010	< .001	0.268	0.012	< .001
φ : change in effort				0.175	0.010	< .001	0.127	0.012	< .001
ω_2: init.eff × pos2							0.040	0.011	< .001
ω_3: init.eff × pos3							0.024	0.011	0.030
ω_4: init.eff × pos4							0.057	0.011	< .001
ω_5: init.eff × pos5							0.058	0.011	< .001
ω_6: init.eff × pos6							0.044	0.011	< .001
ψ_2: change.eff × pos2							0.038	0.011	< .001
ψ_3: change.eff × pos3							0.051	0.011	< .001
ψ_4: change.eff × pos4							0.082	0.011	< .001
ψ_5: change.eff × pos5							0.086	0.011	< .001
ψ_6: change.eff × pos6							0.104	0.011	< .001
Random effects	Var	SD		Var	SD		Var	SD	
σ_{11}	0.932	0.965		0.854	0.924		0.844	0.918	
σ_{22}	1.056	1.028		0.934	0.966		0.934	0.966	
σ_{33}	1.089	1.043		0.974	0.987		0.972	0.986	
σ_{44}	1.130	1.063		0.977	0.988		0.980	0.990	
σ_{55}	1.213	1.101		1.058	1.029		1.061	1.030	
σ_{66}	1.213	1.101		1.060	1.029		1.061	1.030	
σ_β	1.539	1.240		1.538	1.240		1.539	1.240	

Correlation matrix of random effects

	σ_1	σ_2	σ_3	σ_4	σ_5	σ_1	σ_2	σ_3	σ_4	σ_5	σ_1	σ_2	σ_3	σ_4	σ_5
σ_2	0.93					0.92					0.92				
σ_3	0.90	0.94				0.89	0.94				0.89	0.94			
σ_4	0.92	0.95	0.94			0.91	0.95	0.93			0.91	0.95	0.93		
σ_5	0.92	0.92	0.93	0.96		0.91	0.91	0.92	0.95		0.91	0.91	0.92	0.95	
σ_6	0.89	0.90	0.92	0.93	0.95	0.87	0.89	0.90	0.92	0.94	0.88	0.89	0.91	0.92	0.94

Model Fit

AIC	739422	738351	738223
BIC	739742	738694	738681
deviance	739366	738291	738143

Note. Fixed effects terms (δ_2 - δ_6) are effects of positions (see Equation 1.1)
Fixed effects terms ρ and φ are effects of the initial effort and change in effort (see Equation 2)
Fixed effects terms (ω_2 - ω_6) are interaction effects of position and initial effort (see Equation 3)
Fixed effects terms (ψ_2 - ψ_6) are interaction effects of position and change in effort (see Equation 3)
Random effect terms (σ_{11} - σ_{66}) are distribution parameters of θ_i which is multivariate normally distributed (see Equation 1.2)
Random effect term σ_β is the random effect of items: $\beta_i \sim N\left(0, \sigma_\beta^2\right)$.

* $p < .05$; ** $p < .01$; *** $p < .001$.

that the effect for position 4 and position 5 is smaller than the effect for position 3. This reflects the regeneration during the intermission after half of the booklet.

Considering the random effects, and especially the estimates of the person-effects matrix $\boldsymbol{\Sigma}$, two aspects should be noted: First the ability variances (σ_{pp}) are not constant but increase with increasing position. In simple terms: When measured at block position 6, the person variance is higher than if measured at position 1. Second, the correlations of random effects are lower than 1. This indicates that not all examinees are influenced by position effects in the same magnitude, i.e. person effects indeed vary across positions. If we measured examinees' abilities two times, one time only using items at position 1, and another time only using items at position 2, both ability scores would not be correlated perfectly. The correlation declines further if we compare ability estimates of position 1 vs. position 3, for example. More specifically, if the position effect is high, the correlation is comparatively low. Therefore, it is plausible to assume that position effects are moderated by some other variables on the person side.

In model 2, the factor scores of the initial effort and the change in effort from the latent linear growth model are used as additional predictors. Both predictors yielded positive effects. When the initial test-taking effort increases by 1 standard deviation, the logit for correctly answering an item increases by 0.299 on average. This is a replication of previous results which indicate a positive relation between motivation and test performance. To illustrate the impact of this effect, we consider a hypothetical person with ability 0 who invests average effort to solve a hypothetical item with difficulty 0. The correct response probability is $P(X_{ji}=1)=\text{logit}^{-1}(0)=.5$. A second hypothetical person with ability 0 may invest effort of about one standard deviation above average to solve the same item likewise. The correct response probability then is $P(X_{ji}=1)=\text{logit}^{-1}(0+0.299)=.57$.

Regarding the change in test-taking effort, if it is about one standard deviation above the average change of –0.76, the logit for correctly answering an item increases by 0.175 on average.

Model 3 is intended to answer the third research question. In addition, two-way interactions between position effects and both initial effort and change in effort are assumed. With one exception all interaction terms are significant with $p < .001$. The effect sizes, however, are more substantial for ψ_p, which refers to the interaction of position and change in effort. It is indeed the change in effort which essentially moderates item position effects, especially towards the end of the test, i.e. at position four, five and six. To interpret the interaction terms, consider for example the main effect $\delta_6 = -.382$. Item difficulty at block position 6 is on average .382 logits higher than the item difficulty at block position 1. For persons whose change

in effort is one standard deviation above the mean change of -0.76, item difficulty at block position 6 is on average only $.382 - .104 = .278$ logits higher than the item difficulty at block position 1.

Tabe B3 also includes the model fit for all three GLMM. The likelihood ratio test of model 3 vs. model 2 yields that model 3 provides a better fit to the data: $\chi^2 = 147.78$; $df = 10$; $p < .001$.

Discussion

The occurrence of position effects in large-scale assessments of student achievement is a well documented phenomenon. However, research exploring why these position effects occur is lacking. We began to fill this gap by investigating the interdependence of position effects and test-taking effort. To that end, we estimated four consecutive models. The purpose of the first model was to examine the change of test-taking effort over the course of the test. Results of this latent linear growth model yielded that the test-taking effort diminishes considerably during the 120 minutes test. Expressed in standardized coefficients, this decline was 0.76, i.e. at the end of the test (t3), the test-taking effort was about $2 \times 0.76 = 1.52$ standard deviations below the initial test-taking effort.

Person estimates of initial effort and change in effort were then used in three generalized linear mixed models (GLMM) that contain the item responses (correct vs. incorrect) as the dependent variable. All GLMMs allow for multidimensionality on the person side. The first GLMM indicated that persons are differently affected by position effects. Hence, it is plausible to search for variables on the person side which may moderate position effects. We hypothesized that the current test-taking effort is such a moderator variable and added this variable in the second GLMM. Indeed, both initial effort and change in effort were identified as significantly contributing to the probability of solving an item. Finally, interaction terms of positions and initial effort and positions and change in effort were incorporated respectively into the third GLMM. As expected, position effects are more pronounced for persons with lower initial effort and for persons whose test-taking effort declines more rapidly.

These results are relevant for at least three issues in educational measurement: Linking and equating, bias, and validity. First, position effects induced by a fluctuating test-taking effort might be a threat to the item parameter invariance assumption which is central in any linking procedure based on common items. Consequently, examinees' test-taking effort should preferably be kept constant—most desirably at a high level—across studies. Strategies worthwhile

considering might be giving feedback on test results, grading (i.e. consequently converting the test from low to high stakes), and performance-contingent financial rewards. However, Baumert and Demmrich (2001) compared these three alternative approaches to improving the stakes of a mathematical literacy test and found no substantial effects with respect to intended and invested effort or to test performance. Furthermore, the authors concluded that the control condition which merely accentuated the societal utility value of the test is already adequate. In contrast, Marsh (1984) found that performance-contingent rewards, the grading of students, or the evaluation of teachers can lead to an increase in test motivation and performance. However, the present study yielded that examinees substantially differ in the initial test-taking effort. Even if the motivation on average might be enhanced by raising the stakes of testing, the individual differences in test-taking effort persist. There seem to be no panacea to guarantee consistently high effort across examinees during the course of the test.

Secondly, we know from simulation studies that item parameter estimates might be biased due to position effects (Meyers et al., 2009) and that person parameter estimates might be biased due to unmotivated students in low-stakes assessments (Eklöf, 2010). The practical consequences if both occur simultaneously and interdependently are unknown. The present study provides exemplary results which may allow for the further development of simulation designs for examining bias in large-scale assessments due to effects of item position and test-taking effort. In general, simulation studies seldom consider multiple context effects together. However, in practice, researchers are usually confronted with a multitude of context effects that occur simultaneously and are—in worst case—mutually reinforcing.

Thirdly, our results question the validity of the measurement. We see from the first GLMM (model 1), that a persons' competence estimates are not stable across positions. Broadly speaking, what the test measures changes during the test across persons. Persons are affected by position effects in a different way, and these differences are associated with a persons' initial effort and a persons' change in effort. Hence, motivation is blended into the targeted competence measurement. Although this is a known problem in low-stakes tests, our results show that this alteration also varies across persons. Thus, for each person the test measures a somewhat different construct. This actually renders person competence estimates noncomparable, which is a serious threat to validity. The extent of this problem should be investigated in further studies. It is particularly interesting how severely such effects practically affect test scores in commonly used IRT models like Rasch or 2PL/3PL models that ignore motivational context effects.

Several limitations have to be mentioned concerning this study. First, test-taking effort was evaluated via self-reported measures. Though Swerdzewski, Harmes, and Finney (2011) have shown that these are also valid indicators for student motivation in low-stakes tests, Wise and Smith (2011) point out that it is unclear how truthful examinees will indicate their effort. Indeed, a very low test-taking effort may impair the reliability of the scale itself, especially if examinees are requested to complete the effort questionnaire three times during the course of the test. In preparation of the analyses, we had to exclude 1,208 of 47,739 examinees who used the same option for all items of all motivation scales—even on items with reversed polarity.

Furthermore, models from the GLMM framework are always linear in the parameters. This feature allows, for instance, the estimation of the Rasch (1PL) model and extended Rasch-based models like the ones we have proposed. Other IRT models like 2PL/3PL are not supported. Hence, questions as to whether the item discrimination parameter is affected by the item's position or the person's effort cannot be answered. For this purpose, generalized non-linear mixed models (GNLMM) which are, for example, implemented in the NLMIXED procedure in SAS need to be chosen (Debeer & Janssen, 2013).

To conclude, item position effects and persons' test-taking effort do not only exert effects on the responses in large-scale assessments independently, they also interact. This might have serious negative consequences for the reliability and validity of parameter estimates in popular IRT models. We recommend keeping this problem in mind when analyzing data from large-scale assessments and taking preventive measures (e.g., optimizing test-effort) when designing and conducting low-stakes tests.

References

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*(4), 408-426.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.0-6): URL http://CRAN.R-project.org/package=lme4.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441-462.

Brennan, R. L. (1992). The Context of Context Effects. *Applied Measurement in Education, 5*(3), 225-264.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item

response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*(3), 225-244. doi: 10.1177/014662168701100302

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559. doi: 10.1007/s11336-008-9092-x

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1-28.

De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models* (pp. 3-42). New York: Springer.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling*. Mahwah: Erlbaum.

Eklöf, H. (2010). *Student Motivation and Effort in the Swedish TIMSS Advanced Field Study*. Paper presented at the IEA International Research Conference.

Freund, P. A., Kuhn, J.-T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences, 51*, 629-634.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph, 3*, 125-156.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*(3), 379-390.

Hecht, M., Roppelt, A., & Siegle, T. (2013). Testdesign und Auswertung des Ländervergleichs. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (pp. 391-402). Münster, New York, München, Berlin: Waxmann.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M.

(2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*(3), 391-402.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17*(6), 497-509.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models* (pp. 189-212). New York: Springer.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

Kolen, M. J., & Brennan, R. L. (2004). *Testing equating, scaling, and linking: Methods and practice*. New York: Springer.

Marsh, H. W. (1984). Experimental manipulations of university student motivation and their effects on examination performance. *British Journal of Educational Psychology, 54*(2), 206-213.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38-60. doi: 10.1080/08957340802558342

Muthén, L. K., & Muthén, B., O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

OECD. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris, France: OECD.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster, New York, München, Berlin: Waxmann.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific differences. *Large-scale Assessments in Education, 2*(5), 2-17.

R Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM

items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science Quarterly, 51*(1), 47-64.

Swerdzewski, P. J., Harmes, C. J., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*, 162-188.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilseon (Eds.), *Explanatory Item Response Models* (pp. 289-316). New York: Springer.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for two testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*(7), 535-548. doi: 10.1177/0146621614534955

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assesment, 10*(1), 1-17.

Wise, S. L., & Smith, L. F. (2011). A Model of Examinee test-Taking Effort. In J. A. Bovaird, K. F. Geisinger & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139-153). Washington, DC: American Psychological Association.

C. Anhang C, Beitrag 3: Nested Multiple Imputation in Large-Scale Assessments

Dieser Beitrag ist in der Zeitschrift *Large-Scale Assessments in Education* erschienen. Die Referenz lautet:

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T. & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-Scale Assessments in Education, 2*(9), 1-18.

Der Link für den Download des Beitrags ist:
http://link.springer.com/article/10.1186%2Fs40536-014-0009-0